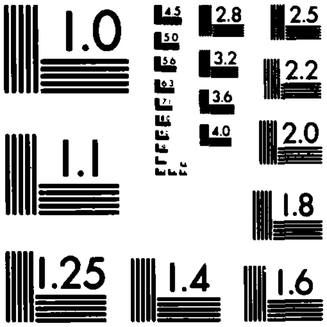
AD-A134 081 A THEORY OF CATEGORIZATION BASED ON DISTRIBUTED MEMORY STORAGE(U) BROWN UNIV PROVIDENCE RI CENTER FOR NEURAL SCIENCE A G KNAPP ET AL. 14 OCT 83 9 N00014-81-K-0136 1/1

UNCLASSIFIED F/G 6/4 NL

END



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(12)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
#9

2. GOVT ACCESSION NO.
AD-A134081

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
A THEORY OF CATEGORIZATION BASED ON DISTRIBUTED MEMORY STORAGE

5. TYPE OF REPORT & PERIOD COVERED
TECHNICAL REPORT

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Andrew G. Knapp
and James A. Anderson

8. CONTRACT OR GRANT NUMBER(s)
N00014-81-K-0136

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Center for Neural Science
Brown University
Providence, Rhode Island 02912

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR 201-484

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training REsearch Program
Office of Naval Research (Code 442PT)
Arlington, Virginia 22217

12. REPORT DATE
October 14, 1983

13. NUMBER OF PAGES
51

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited. Publication in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
Submitted for publication to the Journal of Experimental Psychology: Human Learning and Memory.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Concepts
examplars
neural models
neural nets
prototypes

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
As an alternative to probabalistic and examplar models of categorization, we develop a model based on the assumption of distributed memory storage. Subjects in two experiments performed tasks related to the categorization of random dot patterns. First, the perceived similarity was measured between two such dot patterns, one a distortion of the other. Second, groups of examplar patterns derived from a category prototype were classified together in a category learning task. When the number of examplars was small, new dot patterns were classified according to their similarity to learned

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

exemplars; when the number was large, accuracy depended on a dot pattern's similarity to the prototype pattern. The distributed memory model is used to explain a number of aspects of the experimental findings. Detailed computer simulations are described for the similarity, categorization, and prototype enhancement results.

A Theory of Categorization
Based on Distributed Memory Storage*

Andrew G. Knapp
Massachusetts Institute of Technology

and

James A. Anderson
Brown University

Running Title: A Theory of Categorization

*This work was supported in part by a grant from the National Science Foundation, Memory and Cognitive Processes Section (Grants BNS-79-23900 and BNS-82-14728) and by the United States Office of Naval Research under Contract N00014-81-K-0136. Computer facilities were provided by the Center for Cognitive Science, Brown University. Our thanks to Richard Millward, Director of the Center. We also thank Richard Shiffrin for valuable criticisms and comments.

Abstract

As an alternative to probabalistic and exemplar models of categorization, we develop a model based on the assumption of distributed memory storage. Subjects in two experiments performed tasks related to the categorization of random dot patterns. First, the perceived similarity was measured between two such dot patterns, one a distortion of the other. Second, groups of exemplar patterns derived from a category prototype were classified together in a category learning task. When the number of exemplars was small, new dot patterns were classified according to their similarity to learned exemplars; when the number was large, accuracy depended on a dot pattern's similarity to the prototype pattern. The distributed memory model is used to explain a number of aspects of the experimental findings. Detailed computer simulations are described for the similarity, categorization, and prototype enhancement results.

Introduction

Because events in the world rarely repeat themselves exactly, organisms must possess efficient procedures for relating new information to what has been learned in the past. In particular, the problem faced by a behaving creature is often to reduce the complex and varied inputs it receives into a smaller number of equivalence classes (for example, friend, foe, or food) and to classify correctly novel inputs that are similar to, but not identical with, what has been encountered previously.

Categorization -- the ability to organize information into equivalence classes -- has fundamental importance for any organism that relies on learning for survival. The purpose of this article is to present a theory of memory in which, under certain conditions, categories are formed automatically by the act of storage. We will show that the proposed storage and retrieval scheme captures some of the phenomena observed when human subjects categorize in the laboratory.

Our approach derives from a view of memory that regards information as being stored in a distributed fashion, with separate traces losing their individuality in storage. In this, we differ from many competing models of categorization, and a brief discussion of these differences is warranted.

Models of Categorization

Recently, a great deal of effort has been directed toward the study of human categorizaton (see Mervis & Rosch, 1981, and Smith & Medin, 1982, for recent reviews). According to the formulation of Smith & Medin (1982), models of categorization fall broadly into two families, depending on how categories are assumed to be represented in memory.

Probabilistic models assume that the representation of a category consists of a unitary description of valid category members. However, not all properties of the description are necessarily true of all category members, so that category membership is actually a continuous rather than a two-valued function. This class of models includes the spreading activation model of Collins & Loftus (1975), the feature comparison model (Smith, Shoben, & Rips, 1974) and several models that represent categories as points in a multidimensional space.

One set of probabilistic models we shall be especially concerned with are those that represent a concept by an average (sometimes called a prototype) of category instances. A body of empirical evidence has been used to support the idea that, in some tasks, subjects abstract a prototype from items classified together during learning, and that novel items are classified according to the prototype they most resemble. This evidence includes the finding that category prototypes are sometimes classified more accurately than other category members, including the exemplars actualy seen during learning (Franks & Bransford, 1971; Posner & Keele, 1970, Strange, Keeney, Kessel & Jenkins, 1970; also see Robbins, Barresi, Compton, Furst, Russo & Smith, 1978), suggesting that the prototype is a main constituent in the category's mental representation.

Exemplar models make up the other main family of categorization models. According to this view, no single description of a category exists. Rather, an aggregate of separate descriptions of some or all category members serves to represent the category (Brooks, 1978; Nelson, 1974). According to this view, stimuli are categorized according to the number of stored exemplars they retrieve. These proposals include the proximity and best-examples models (Reed, 1972) as well as the context model of Medin & Schaffer (1978). Proponents of exemplar models have pointed out that many of the findings taken to support prototype abstraction can be equally well explained if only the learned instances are assumed to be stored in memory (see, for example, Hintzman & Ludlam, 1980).

Although they lead to quite different approaches to the problem of categorization, the memory representations postulated by both probabilistic and exemplar models are similar in an important respect. Both views assume the descriptions representing learned categories to be stored separately from one another. In the case of exemplar models, the descriptions of individual category members which combine to make up a category's representation are likewise assumed to be separate, identifiable entities. The idea that items stored in memory reside in unique 'locations' or form distinct 'traces' is an implicit assumption common to both families of models. Rejecting the assumption of separate storage leads to a third class of models, which we will call distributed memory models of categorization.

The fundamental assumption behind the distributed memory approach is that remembered items share many or all of the same storage elements, so that one cannot properly point to a single memory 'trace.' Previous theoretical work has demonstrated that information can be both stored and retrieved without assuming separate storage of individual items (for reviews see Kohonen, 1977; Hinton & Anderson, 1982, Anderson & Hinton, 1982). In addition, distributed-memory models have been applied to a variety of cognitive processes, including associative learning (Anderson, 1983; Murdock, 1983; Eich, 1982), list learning tasks (Anderson, 1973; Anderson, 1977), as well as categorical perception, distinctive feature analysis and probability learning (Anderson, Silverstein, Ritz & Jones, 1977). A model with important similarities, in that it uses a parallel system with diffuse connectivity, the 'interactive activation' model for word recognition has been described and successfully applied to much experimental data. (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982).

It is worth emphasizing that the assumption of a distributed memory does not strictly preclude the more familiar models of categorization. It is perfectly possible to implement probabilistic or exemplar model 'software' on distributed memory 'hardware.' However, the converse is not always true: the behavior of some distributed memory models can be mimicked only with ad hoc assumptions or at great computational expense if separate storage is assumed. We will illustrate this point by developing a distributed memory model that acts as a simple categorizer. Then, we will compare the results of two experiments pertaining to categorization with the predictions of the model. Where appropriate, we will contrast the proposed model with both probabilistic and exemplar models, but our main intent is to demonstrate the viability of a distributed memory approach to categorization.

A Distributed Memory Categorizer: Introduction.

The model to be presented was inspired by speculation about how associative learning might occur in the nervous system, but we will develop it here without making any claims about physical realization. A review of some neuroscientific evidence bearing on the model is available (Levy, Anderson & Lehmkuhle, 1984). More detailed presentation of the material in the following section, with numerical examples is available (Anderson, Silverstein, Ritz & Jones, 1977; Anderson & Hinton, 1981).

We begin by assuming that a large number of richly interconnected but rather simple elements participate in the storage of information. This is the basic distributed memory assumption. We refer to these elements as 'neurons' and to their connections as 'synapses', while remaining agnostic about their possible realization in real nervous systems.

In our idealized scheme, each neuron has an 'activity' that depends on the synaptic inputs it receives from other neurons, where a synaptic input is defined as the activity of the input neuron multiplied by a weighting factor that we will call the 'strength' of the synapse. In particular, we assume that neurons behave as linear integrators: a neuron's activity is simply the weighted sum of its synaptic inputs. Thus, if a neuron receives inputs from three other neurons whose activities are 20, -10, and 2, with synaptic strengths 0.5, 0.2 and -1.0 respectively, its activity will be

$$(20)(0.5) + (-10)(0.2) + (2)(-1.0) = 6$$

Note that both neuronal activities and synaptic strengths can take on negative as well as positive values.

Information is represented in such a system by the pattern of activty across a large number of neurons. Formally, we denote these activity patterns by $\underline{N}$-component vectors, where each element in the vector is the activity of a single neuron and $\underline{N}$ is the number of neurons. Now suppose there are two such sets of $\underline{N}$ neurons, alpha and beta, connected so that every neuron in beta receives an input from every neuron in alpha as illustrated in Figure 1.

Figure 1 about here

We can conveniently represent the *N* squared synaptic strengths by an *N* by *N* connectivity matrix A, where each entry A (*i*,*j*) is the strength of the synapse between neuron *i* in alpha and neuron *j* in beta. In the absence of other inputs, the activity of each neuron in beta is thus completely determined by the activities of the neurons in alpha and by the synaptic strengths.

In vector notation, this relationship is

$$(\text{pattern in beta}) = A\ f.$$

Information can be stored in such a system by modifying the synaptic strengths as follows. Suppose that all the strengths are initially zero, and that activity pattern f occurs in alpha and pattern g simultaneously occurs in beta. Here f might denote a stimulus and g a response that has just been rewarded. We assume that in such a situation the synaptic strengths are able to change according to the rule

$$A\ (i,j) \propto f\ (i)\ g\ (j),$$

That is, the strength of each synapse is incremented proportionally to the product of the pre- and postsynaptic neurons (cf. Hebb, 1949).

For illustration, suppose that the proportionality constant in the above learning equation is one and the vector f is normalized so that the length of f is set equal to one and A is initialized to 0. (i.e. the inner product, $f^T f$ is one and A is all zeros). The resulting connectivity matrix becomes

$$A = g\ f^T$$

Now suppose that after the synapses comprising A have been modified, pattern f again occurs in alpha. By the above,

$$\begin{aligned}\text{(pattern in beta)} &= A\, f \\ &= g\, f^T f \\ &= g\end{aligned}$$

Therefore, the result of this form of synaptic modification is that subsequent occurences of f in alpha give rise to g in beta -- the two patterns have become associated.

In general, such a simple system can associate as many as *N* pairs of patterns, (f_1, g_1), (f_2, g_2), (f_3, g_3) . . . (f_N, g_N) though the practical capacity is less. Each association increments the matrix according to the above rule, so the final matrix becomes

$$A = \sum_{i=1}^{n} g_i f_i^T$$

Although the *n* associations are spread over and mixed together at the same *N* synapses, information may not have been lost.

When one of the f's, say f_i, occurs in alpha then

$$\begin{aligned}\text{(pattern in beta)} &= A\, f_i \\ &= g_i f_i^T f_i + \sum_{i \neq j} g_j (\, f_j^T f_i).\end{aligned}$$

Consider the special case where the simulus set (the f's) are orthogonal, that is, the f's are at right angles to each other and their inner product $f_i^T f_j = 0$ when *i* $\neq$ *j* and one when *i* = *j*. Then the

$$\text{(pattern in beta)} = g$$

since all the inner products are zero. This is a more useful approximation than it seems, because if *N* is large and the f's have statistically independent components, they will be close to orthogonal.

Categorization

The model can function as a simple categorizer by making one additional assumption. *Let us make the fundamental coding assumption that the activity patterns representing similar stimuli are themselves similar, that is, their state vectors are correlated*. This means the inner product between the two patterns is not small.

Now consider the case described above where the model has made the association (f, g). Let us restrict our attention to the magnitude of the vector in beta that results when various patterns occur in alpha. We have just shown that when f occurs in alpha, g occurs in beta. When a new pattern f' occurs in alpha, then

$$\text{(pattern in beta)} = g\, f^T f'$$

If f and f' are uncorrelated, their inner product $f^T f'$ is small. If f is similar to f' then the inner product will be large. The model responds to input patterns based on similarity to f. Patterns similar to f give strong responses (as measured by the length of the output vector) while dissimilar patterns produce weak responses. Thus, the nature of the learning assumption gives us an automatic generalization mechanism. Furthermore, this formulation suggests that the perceived similarity of two stimuli should be systematically related to the inner product $f^T f'$. We test this prediction in Experiment 1. If the network has learned several associations, it can categorize novel input patterns according to their similarity to the patterns already encountered.

Multiple Member Categories

This is a rather limited form of categorization, however, because each category has only one member. We will now apply the distributed memory model to a more realistic situation where a category contains many similar items. Here, an entire set of similar activity patterns (representing the category members) becomes associated with the same response, for example, the category name. It is convenient to discuss such a set of vectors with respect to their mean. The mean is taken over all learned members of the category; if the categorizing system does not see all the members of the category but some subset of them, the interesting behaviors discussed next can appear.

Specifically consider a set of correlated vectors f_1 . . . f_n with mean p. Each individual vector in the set can be written as the sum of the mean vector and an additional noise vector, d_i, representing the deviation from the mean, that is,

$$f_i = p + d_i .$$

When these n patterns occur in alpha and are all associated with th● same response, g, in beta, the final connectivity matrix will be

$$\begin{aligned} A &= \sum_{i=1}^{n} g\, f_i^T \\ &= g \sum_{i=1}^{n} (p^T + d_i^T) \\ &= \underline{n}\, g\, p^T + g \sum_{i=1}^{n} d^T \end{aligned}$$

The term containing the sum of the noise vectors (the d_i) is particularly important. Suppose that this term is relatively very small, as would happen if the system learned many randomly chosen members of the category. In that case, the connectivity matrix reduces to

$$A = \underline{n}\ g\ p^T .$$

The system behaves as if it had repeatedly learned only one pattern, p, the mean of the correlated set of vectors it was actually exposed to. Under these conditions, the simple association model can extract a 'noisy signal', just like an average response computer. In this respect the distributed memory model behaves like a prototype model, because the most powerful response will be to the pattern p, which may never, in fact, have been seen.

However if the sum of the d terms is not small, as might happen if the system only sees a few of patterns from the set, the response of the model will depend on the sum of the similarities between the novel input and each of the learned patterns, that is, the system behaves more like an exemplar model. If the perturbing noise vectors are of roughly constant size then the number of patterns in the correlated set of inputs will be the primary factor determining whether the distributed memory model behaves more like a prototype model or an exemplar model. This is the topic of Experiment 2.

Finally, we need to consider what happens when members of more than one category occur in alpha. Suppose the system learns items drawn from three categories with means of p_1, p_2, and p_3 respectively. If g_1, g_2, and $g_3$are the responses associated with the three categories, then

$$A = g_1 s_1 + g_2 s_2 + g_2 s_3$$

where each sum s is once again of the form

$$s_i = \underline{n}\ p_i + \sum_{j=1}^{k} d_j .$$

To determine which response, g_1, g_2, or g_3, most closely matches the model's output when presented with an input f it is sufficient to examine the inner products,

$$s_1 f\ ,\ s_2 f\ ,\ \text{and}\ s_3 f\ ,$$

and choose the largest. That is, the item will be classified according to its similarity (measured, as before, by the inner product) to one of the stored sums. Due to superposition (this is a linear system) the actual response pattern will be a weighted sum of the three responses. If the inputs are reasonably well separated (i.e. the inner product between different inputs is small) the distortion of the appropriate output will also be small. (This seems to be the case in the our experiments, based on subject's responses to purely random stimuli, as we shall describe.) If felt necessary for theoretical adequacy, however, we can invoke a distributed non-linear feedback model related to the model just presented, which can correct distortions and exactly reproduce the correct output response. This model has been described elsewhere. (Anderson et al., 1977). It, too, calculates similarity to the stored sums.

Random Dot Pattern Stimuli

In order to perform experiments that could be plausibly related to the distributed memory model, it was necessary to use stimuli whose relations to one another can be readily quantified. In addition, we wanted to avoid well established natural categories such as common objects and facaes, because the model's behavior depends in unpredictable ways on what stimuli have been encountered in the past. These considerations led us to adopt the artificial categories first conceived by Posner and Keele (1968, 1970).

These stimuli consist of nonmeaningful arrangements of dots. In a typical experiment, several readily distinguishable patterns are produced by randomly distributing nine dots within a display area. These original patterns are called prototypes. A family of distortions of each prototype can then be generated by moving dots random distances in random directions according to various rules. A single distortion of a prototype is calleo, in this literature, an exemplar. By manipulating the motion of the dots, category exemplars can be either grossly distorted versions or only slight variations of the prototype. Each prototype and its progeny constitute an artificial category. Like many natural categories, dot pattern categories are ill defined (Neisser, 1967) because any pattern can be transformed into another by an appropriate distortion.

Because all the experiments used the same basic materials, the methods used to generate and present stimuli will be described in detail at this point. Specific experimental procedures will be described as each experiment is introduced.

General Method

The subjects were all paid volunteers from the Brown University Psychology Department's pool of students and staff. Subjects were assigned to experiments by order of appearance in the laboratory and were tested either individually (Experiment 1) or in groups of three (Experiment 2).

The studies took place in Brown University's Human Learning Laboratory (Millward, Aikin & Wickens, 1972). The experiments were controlled online by a Digital Equipment Corporation PDP-8/e minicomputer which generated all the stimuli, determined display order and timing, collected subject's responses and reaction times, and gave subjects feedback about their responses. The dot pattern stimuli appeared on Tektronix 502 oscilloscopes. The display area of each oscilloscope screen measured 10 by 10 cm and was divided into a 512 by 512 unit grid, this grain being determined by the digital-to-analog converters used to transmit the stimuli from the computer to the oscilloscopes. The intensity of the oscilloscope beam was adjusted so that the stimulus dots were as small as possible while remaining clearly visible.

The subjects sat in dimly lit booths. The display screens were viewed through openings in the front wall of each booth, approximately 75 cm from the subjects; the dot patterns thus subtended about six degrees of visual angle. Subjects indicated their responses by pressing keys on teletype keyboards located below and in front of the display windows. A row of computer operated lights mounted on the keyboards provided feedback when required.

All the experiments involved two types of stimuli: prototypes and exemplars. Each prototype pattern was created by randomly placing nine dots within a 300 by 300 unit grid centered in the 512 by 512 unit display area. Exemplars of a prototype were constructed by displacing the prototype dots short distances. For each dot, the direction of motion was chosen at random (i.e. from a uniform distribution on the range zero to 360 degrees. The distance moved by a dot was determined by a probability distribution specified by the experimenters.

Figure 2 shows a prototype and five exemplars at increasing levels of distortion. Each exemplar is the result of moving the prototype dots distances drawn from a normal distribution. The mean of the distribution (in display screen units) is given above each exemplar. The standard deviation of the distribution was one third of the mean in all cases. Because the prototypes were generated toward the middle of the display screen, the dots had ample room to move. The computer halted at the border any dots which attempted to stray beyond the confines of the display area, but such events were rare, even for the largest displacments used.

Figure 2 about here

The significance level for all statistical tests was 0.05.

Experiment 1. Similarity Between Two Dot Patterns

It is apparent from Figure 2 that the resemblance between an exemplar and its prototype decreases as the dot displacements used to create the exemplar increase. As described above, our distributed memory model predicts a general form for this relationship. According to the model, the perceived similarity between two stimuli is related to the inner product **f f'**, where **f** and **f'** are the distributed activity patterns representing the stimuli. Testing this prediction requires (a) empirical mesurements of perceived similarity as a function of distortion, and (b) a detailed formulation of the distributed representations of stimuli, which until this point have been treated only as abstract vectors. In Experiment 1, therefore, subjects judged the similarity of prototype-exemplar pairs at several levels of distortion. The dot-displacement probability distributions were systematically varied in an effort to determine which parameters of dot displacement influence perceived similarity. The results of this experiment lead us to propose a specific distributed representation for dot-pattern stimuli. Incorporating this representational scheme into the model allows us to predict subjects' similarity judgments quantitatively.

The present experiment extends Posner's (1964; Posner, Goldsmith & Welton, 1967; see also White, 1962) investigations of the determinants of similarity between dot patterns. Those studies employed a variety of distortion rules, but found that subjective similarity seemed to depend only on the average distance moved by the prototype dots during the creation of an exemplar. Subsequent investigators (Barresi, Robins, & Shain, 1975; Homa, Cross,Cornell, Goldman, & Shwartz, 1973) have used still other distortion rules but have continued to rely on this average distance metric (or on an equivalent metric cast in the language of information theory) as an objective index of similarity.

It is a property of the average distance measure that vastly different sets of individual dot displacements can give rise to the same mean displacement. In particular, all the individual dots can move roughly the same distance (low variance about the average) or some dots can move a great deal, others not at all (high variance, cf. Barresi et al., 1975). However, the effect of variance on perceived similarity has not been systematically investigated. In the present experiment, similarity ratings were obtained for pairs of patterns which bore either a high- or a low-variance relation to one another. If average displacement is an adequate similarity metric, the variance manipulation should have no effect on similarity judgments.

Method

Subjects. 10 members of the Brown University community received $3.00 for participating in a single hour-long session. Subjects were unfamiliar with the stimuli and unaware of the purpose of the experiment. All subjects were tested individually.

Stimuli and design. Subjects viewed pairs of sequentially presented dot patterns. The first member of each pair was a prototype (generated as described previously), and the second was an exemplar derived from that prototype. Clearly, the terms 'prototype' and 'exemplar' are somewhat arbitrary in this context, because a new prototype was generated for each trial and no prototype was repeated. A new set of stimuli was created for each subject.

We manipulated the distortion used to create the exemplar patterns as follows: Five levels of fixed distance distortion were obtained by moving all the dots in the prototype pattern the same distance -- 15, 30, 45, 60, or 90 units -- in random directions. Corresponding levels of variable distance distortions were created by moving five of the prototype dots a shorter distance than these same values, and moving four dots a longer distance. The actual displacements used are shown in Table 1. This procedure ensured that each level of variable-distance distortion involved a smaller average displacement but a larger displacement variance than the corresponding level of fixed distance distortion. For example, in generating exemplars at the third level of variable distance distortion, the prototype dots moved an average of 40.7 units (versus 45.0 units for the corresponding level of fixed distance distortion), although each individual dot could move as much as 90 units or as little as 0 units (see Table 1).

Table 1 about here

Each subject rated 20 different prototype-exemplar pairs at each level of fixed distance and variable distance distortion, and 20 patterns in which the two patterns were identical. Trials were divided into two blocks of 100 separated by a short rest period. Within a trial block, there were 10 pairs at each distortion level, ordered randomly for each subject.

Together with subjects' similarity ratings, two measures were recorded for each pattern pair: (a) the average distance moved by the dots, defined as the sum of the individual dot displacements divided by nine (a displacement being the distance from the dot's initial screen position in the prototype pattern to its final position in the exemplar pattern), and (b) the root-mean-square (RMS) distance moved, defined as the square root of the mean of the squared invididual displacements. The latter measure incorporates information not only about the mean of the displacement distances, but also about their variance.

Procedure. Subjects were told they would be shown a series of dot pattern pairs, presented sequentially. They were instructed to rate each pair on an eight point similarity scale -- one meaning highly similar and eight meaning very different --

by pressing the appropriate teletype key. Subjects were not told that any of the patterns would be identical.

The first member of each pair was presented for five seconds, followed by a one second interstimulus interval (ISI), folowed by the second pattern, which remained visible until the subject responded. A two second interval followed each response, during which time a keyboard light indicated that the subject should prepare for the next trial.

At the start of the experimental session, each subject rated 33 practice trials, three at each distortion level in random order, to become familiar with the procedure and with the range of similarities.

Results. Figure 3 plots the mean similarity ratings for each level of fixed distance and variable distance distortion as a function of average dot displacement. The identical patterns received a significantly larger rating than the minimum possible value of 1.0 ($z(200) = 2.02$), while the largest mean rating (6.76 for the most distorted variable distance pairs) was substantially less than the theoretical maximum of 8.0. These findings may reflect a reluctance on the part of the subjects to use extreme ratings.

Figure 3 about here

Nevertheless, the main result of the present experiment is quite clear: variable distance pairs were consistently judged more different than corresponding fixed distance pairs, even though the average distance moved by the dots was always less for variable distance distortions than for corresponding fixed distance distortions. To test this result statistically, each point on the variable distance and fixed distance curves in Figure 3 was compared with the interpolated point below or above it on the other curve in a large sample z-test (one-tailed). Standard deviations of these interpolated similarity ratings were estimated by choosing the standard deviation of the nearest point on the same curve. This procedure is justified, since the variability of the ratings was fairly constant (range of standard deviations: 1.1 - 2.2). By this test, mean similarity ratings for variable distance pairs were significantly larger than for corresponding fixed distance pairs ($z(200) > 1.76$ in all cases).

The results demonstrate that average dot-displacement is an insufficient measure of similarity when some dots move a great deal more than the average and others move much less. A similarity measure weighting large dot movements more than small ones would predict the present data more accurately, since large displacements occur preferentially in the variable distance conditions. One (though by no means the only) such measure is the RMS distance metric described above.

Discussion and Theory

The finding that dot displacement variance influences perceived similarity seems consistent with a study by Barresi et al. (1975) in which the prototype of one dot pattern category was actually a large distortion of the prototype of a second category. Subjects learned to classify exemplars of the categories more quickly when the two prototypes were high variance distortions of each other than when they shared a low variance relation, the mean dot displacement being the same for both prototype pairs. The results of the present experiment suggest that categories derived from high variance prototype pairs are easier to learn because the prototypes are less similar to one another than in the low variance case.

Theory: Similarity. To obtain quantitative similarity predictions from the distributed memory model we must specify a neural coding of the input patterns, that is, we must describe how the state vectors are generated from the physical input pattern. We know from many sources that there is a powerful topographic mapping of visual space onto the surface of the cerebral cortex. Let us consider the coding due to a single dot in visual space. We assume that this will map onto a 'bump' of activity on a hypothetical surface composed of many elementary 'neurons' which represents the neural coding. Since real neurons have receptive fields of varying width, even in a single cortical region, we assume there is a fall off of activity from the central location corresponding to the exact topographic location of the dots: some cells have large receptive fields and respond to a dot even if the center of their fields is far away from the dot location; other receptive fields are much smaller and must be precisely centered on the dot location. Figure 4 shows an exponential decay of activation, that is the activity at a point, $a(\underline{r})$, is given by

$$a(\underline{r}) = \exp(-\underline{r}/\lambda)$$

where λ is the length constant of the exponential and $\underline{r}$ is the distance from the center of the distribution to the point whose activity is to be computed.

Figure 4 about here

Interactions Between Activity Patterns. We are concerned with inner products between two dot patterns. We can compute the inner product between the activity patterns of any two single dots. If our activity patterns are composed of nine dots (as ours were) then the basic linearity of the model lets us then consider all pairs of dots (81 pairs) and add up the resulting inner products giving the overall result. Note that the model does not need to "match up" dots from the two stimuli; any activity patterns that interact will contribute to the inner product.

We investigated a number of different activity patterns due to single dots. The best fit to the data was obtained with the exponential distribution shown in Figure 4. In this case, we are able to obtain a closed form solution to the dot product. Suppose we have two activity patterns due to single dots separated by a distance $\underline{d}$. Let us assume the distribution is continuous, though in reality it is made up of discrete neuron-like elements. For convenience let the length constant equal one, and consider two exponential distributions, $a(\underline{x}, \underline{y})$ and $b(\underline{x}, \underline{y})$ separated by a distance $\underline{d}$ along the $\underline{x}$ axis. We want to evaluate the integral, $I(\underline{d})$, which will only be a function of displacment, $\underline{d}$:

$$a(\underline{x}, \underline{y}) = \exp\left(-\sqrt{(\underline{x} + \underline{d}/2)^2 + \underline{y}^2}\right)$$

$$b(\underline{x}, \underline{y}) = \exp\left(-\sqrt{(\underline{x} - \underline{d}/2)^2 + \underline{y}^2}\right) \text{ and}$$

$$I(\underline{d}) = \int\int a(\underline{x}, \underline{y})\, b(\underline{x}, \underline{y})\, \underline{dx}\, \underline{dy}.$$

Let us normalize the function so that $I(\underline{d}) = 1$ when $\underline{d} = 0$. This integral can be computed exactly and is given by

$$I(\underline{d}) = (1/2)\, \underline{d}^2 K_2(\underline{d})$$

where $K_2(\underline{d})$ is a modified Bessel function of order 2. (See Abramowitz & Stegun, 1964). Figure 5 shows shows $I(\underline{d})$ graphed with the ordinate inverted for comparison with Figure 3. This computed function for a single pair of dots will approximate the more complex situation where nine dots are involved, especially at small displacements.

Figure 5 about here

For connoisseurs of integration, this integral can be done by observing that the loci of constant product are ellipses. The equations are converted to elliptical coordinates (Korn & Korn, 1968) and then integrated with the tables in Gradshteyn & Rhyzik (1959).

Computer Simulations of the Similarity Experiments. Let us first describe the simulations of Experiment 1, the experimental measurements of similarity. Computer simulations allows us to handle these multiple dot experiment exactly. We can generate patterns like those used in the experiments, distort the patterns, compute the inner product, and then compute a measure of goodness of fit of the inner product to the experimental data.

Note that in this simulation, we have 81 pairs of dots (9 dots taken two at a time, one from a pattern and the other from its distortion) between which we must compute inner products. In the computer program, we formed a matrix of distances between dots. Note that the diagonal of the matrix of distances represents distance between the a dot and its displacement in the distortion. If average displacement was small, we would expect most of the contribution to the inner product to be concentrated along the diagonal. As the displacement increased, we should expect more contribution to move off the main diagonal. The displacements used in our experiments had a significant contribution from the off diagonal elements. Using the best fitting length constants computed from the similarity experiments the sum of the inner products due to the off diagonal terms and the sum of the inner products due to the diagonal terms were roughly equal.

The off diagonal elements of the matrix of distances also contain information about the entire pattern, in fact enough to reconstruct the pattern bar rotations and inversions. A criticism is sometimes made of these models that one could present dots one at a time and one would predict the same results as if all nine dots were presented at once, since only the sum of activities is involved. There are two immediate responses to this: First, one pattern of nine dots looks grossly like another pattern of nine dots. Though we are only concerned with differences between patterns, surely the full neural codings contain a significant context which is identical between different patterns: i.e. dot number, the experimental situation, etc. This constant part plays no part in our computations because it is identical for all the dot patterns, but if we make a significant change in experimental situation, it will become important. Second, the off diagonal and diagonal terms together respond to all possible dot-dot interactions in the pattern. Dots presented one at a time would have no off diagonal elements unless some rather arbitrary assumptions (things are summed up in a buffer, etc.) were made which reestablish the pattern nature of the stimulus.

Programs were written in Pascal, which is exceptionally convenient for this kind of simulation. We defined, for example, a Pascal RECORD 'Dot_Pattern' which has all the formal properties of a dot pattern. Copies of these programs are available.

By assumption, magnitude of inner product between activity patterns is directly related to similarity. The only free parameter in the simulation was the length constant.

In the simulations, mathematical representations of the patterns were constructed which were statistically identical to those actually presented. For the similarity simulation, a dot pattern and a distortion were generated and the inner product of the neural codings of the two patterns (using exponential decays) were computed. The experimental and theoretical values were compared. Several measures of goodness of fit were used in the computations. First, the best fitting straight line was computed between predicted similarity and experimental similarity. The length constant which minimized the mean square distance between this line and the experimental data was found. Second, the correlation between predicted and experimental values was computed.

The results of the measures were close to each other. Also, the maxima of the relation between length constant and the measures of goodness of fit was quite broad. There were no critical aspects of the simulation.

Several functions representing falloff of activity due to a single dot were investigated at various times: Gaussians, exponentials, laterally inhibited functions, and others. The best fits were obtained with simple exponentials and this function was used in the simulations and in the figures. Exact shape of the falloff not critical, that is, the pattern of fits obtained was roughly the same for different falloffs, but best fits were numerically not quite as good.

Experimental and theoretical fits are given in Table 2. This table was computed from 100 simulated dot patterns at each level of distortion and is a typical simulation. Values of parameters were those used for best correlation between simulated and experimental values. Best fitting length constant was computed for five different sequences of patterns. Each sequence was generated from a different random number generator seed. The average best fitting length constant for maximizing correlation between computed and experimental similarity was 14.5 screen units, producing an average correlation of 0.97. At this value of length constant the mean square difference between the simulated values and the best fitting line between experimental and simulated values was 0.215. The best fitting length constant for minimizing mean square difference between experimental and computed values was 11.7 screen units (producing a mean square difference of 0.210) and the average correlation was over 0.96.

Table 2 about here

Experiment 2: Number of Exemplars

Experiment 1 demonstrated good agreement between the distributed memory model's predictions and subjects' behavior when asked to rate the similarity between two dot patterns. A particular formulation of the model was developed to fit the similarity model quantitatively. We next sought to test this same model's predictions for a true categorization tasks. As described above, the model represents a category by associating an entire set of correlated activity patterns with the same response. If the activity patterns are of the form

$$f_i = p + d_i$$

where p is the mean of the correlated set and d_i is the deviation of the individual pattern from the mean, and if the d's are of roughly constant magnitude, then the model's behavior depends mainly on the number of activity patterns in the learned set. If the model learns many different category members, it behaves if it has seen the mean pattern p alone whereas if it learns only a few members, those learned patterns will dominate the category's representation.

In Experiment 2, subjects learned three categories containing 1, 6 and 24 members respectively. The learning stimuli were exemplars of one of three prototypes, with the degree of distortion between each exemplar and its prototype held constant. Only the number of learned stimuli varied across the categories. Subjects learned the categories by classifying the exemplars and receiving feedback about their decisions. During this process, the categories were presented with equal frequency, to avoid biasing subject's classifications. Following the learning phase of the experiment, subjects classified (without feedback) a series of dot patterns, including the training stimuli again, new exemplars created from the same category prototypes, and the prototypes themselves.

Method

Subjects. 21 paid subjects from the same pool drawn upon in Experiment 1 participated in a single 20 minute session. To ease data analysis, the subjects were tested in groups of three. Within each group, all three subjects saw the same stimuli in the same random order.

Stimuli and design. The experiment had a learning phase, in which subjects classified category exemplars with feedback, and a testing phase, in which they classified old, new, and prototype patterns without receiving feedback. For each group of subjects, three prototypes (designated A, B, and C) were constructed. Exemplars of the prototype were created by the method described earlier. The dot displacement were drawn from a normal distribution with mean 24 and standard deviation eight; the average RMS distance between prototype and exemplar was 25.2 units. This distance was chosen in pilot studies because it seemed to maximize the experimental prototype enhancement.

The learning stimuli consisted of one exemplar of prototype A displayed 24 times, six exemplars of B displayed four times each and 24 exemplars of C shown only once each, for a total of 72 learning trials, with each category being represented 24 times during learning. In the testing phase, the single learned exemplar from category A was presented eight times, four of the old exemplars from category B were presented twice each, and eight of the old C exemplars were each presented once. In addition, eight newly constructed exemplars of each prototype were presented once each and each prototype was shown eight times. The subjects had encountered neither the prototypes nor the new exemplars during learning. Finally, nine unrelated control patterns (newly generated prototypes) were presented once each for a total of 81 test trials: eight old exemplar trials, eight new exemplar trials, and eight prototype trials for each category, plus the nine control trials.

Procedure. The subjects were told that they would be shown a series of dot patterns, and that their task would be to determine which patterns were to be grouped together under the same response. Subjects were instructed to classify each pattern by pressing one of the three teletype keys. The keys were randomly paired with the categories at the start of the session and these pairings were the same for all subjects within a group. All subjects responded with the forefinger of the dominant hand; between responses they placed the response finger on a spot about 2 cm from the response keys and approximately equidistant from them.

Stimuli were presented one at a time until all three subjects had responded or for a maximum of five seconds. In the learning phase, subjects were urged to respond during the display interval if possible, but not to rush their responses. Each stimulus presentation was followed by a five second feedback interval, during which the correct response for that trial was indicated by illuminating a light above the appropriate response key. In this part of the experiment, subjects were instructed to concentrate primarily on learning the classifications and were encouraged to guess during the initial trials.

In the testing phase, a five second blank ISI followed each stimulus presentation. Subjects were instructed to respond as quickly as possible, based on what they had learned in the first part of the experiment, and were informed that no feedback would be given.

Figure 6 about here

Results

The data from two subject groups had to be discarded due to equipment failure. Of the remaining 15 subjects, one performed below chance level during the learning phase, but these data were included because this subject performed at a high level of accuracy during the testing phase.

The mean number of errors made in the learning phase was 15.1 (range 4 - 43). For categories A, B, and C, these means were 2.8, 6.2, and 6.1 respectively (ranges: 0-16, 0-17, 1-10). Subjects made signficantly fewer errors learning the one exemplar category than the other two categories (*S* = 7.3 on a Friedman test); several subjects made only one error on this category during the entire learning phase. This advantage did not transfer to the testing phase, where the overall classification accuracy was 88% for categories A and B and 91% for category C.

Figure 6 shows the percent correct classification for each test trial combination of stimulus type (old, new, prototype) and number of learned exemplars (1, 6, 24) collapsed across subject groups. A ceiling effect is apparent for the single old exemplar in the one-instance category: all the subjects classified this pattern correctly all the time.

An analysis of variance was performed on the classification data, treating stimulus type and number of learned exemplars as within subjects variables, and group as a between subjects variable. The main effect of Type was significant [*F* (2,20) = 7.21, $\underline{Ms}_e$= 0.21] but the main effect of Number and Group were not (both *F*'s < 1). The Type by Number interaction reached significance [*F* (4,40) = 3.99, $\underline{Ms}_e$= 0.22]; performance on the old exemplars decreased with larger numbers of learned items, while performance on the new and prototype instances increased. The prototypes were always classified more accurately than other novel exemplars, and by about the same margin. No other interactions approached significance. In particular, Block did not interact with either of the within-subjects variables. This finding is reassuring, since it suggests that the present results do not depend critically on peculiarities of the randomly generated stimuli that we used.

Reaction time (RT) data were collected to provide convergent support for conclusions drawn from the classification data. Previous studies using dot pattern stimuli (Homa et al., 1973; Posner & Keele, 1968; see also Omohundro & Homa, 1981) have found that response speed tends to correlated with classification accuracy. The RT data, summarized in Figure 7 generally confirm the pattern of the error data. The latencies indicated that the classification results are not the product of a speed-accuracy tradeoff.

Figure 7 about here

An analysis of variance was performed on the RT's to correct responses, with RT's that exceeded a subject's mean by three standard deviations or more deleted. As in the classification data, the main effect of Type was significant [*F* (2,20) = 10.06, *MS* = 29.90] while the effect of Number was not (*F* < 1). The Type by Number interaction was again significant [*F* (4,40) = 3.30, *MS* = 48.67]. In addition, the main effect of block was significant [*F* (4,10) = 5.15, MS_e= 566.71] indicating that while the specific stimuli used did not affect subject's accuracy, they did affect how long it took to make their classifications. No other interactions reached significance (*F* < 1, in all cases).

Subjects assigned the unrelated control patterns to the one, six and 24 instance categories 5.9%, 31.1%, and 63.0% of the time, respectively. This corroborates the finding of Homa et al., (1973) that subjects tend to assign random patterns to the category containing the largest number of learned instances. The mean RT for the control patterns was 1,556 msec, about 350 msec longer than for the other stimuli. Clearly, these patterns are recognized as not belonging to any of the three learned categories. This finding provides reassurance that prototypes chosen at random differ from one another a good deal, so there is little category overlap in these experiments. This was the subjective impression of the experimenters and seems to be consistent with the data.

Discussion and Theory

The results of Experiment 2 support the qualitative predictions of the distributed memory model. While varying the number of learned exemplars did not substantially affect subjects' overall accuracy in classifying test stimuli, it did affect their relative accuracies for the different types of category instances. Old exemplars were classified more accurately for the smallest category, and a prototype advantage was apparent for the largest category. We have proposed that this result derives from representing a category as the sum of the distributed activity patterns representing the category members. We have also proposed a specific form for the activity patterns representing dot patterns, so we can model Experiment 2 without making any additional assumptions.

Figure 8 about here

Figure 9 about here

The model's behavior is more understandable if we begin by considering the summed representation of a single dot in the nine dot patterns. (Figures 8 and 9) These diagrams may be considered "close-up" views of the category's representation, with similar summation occuring in the regions of the other eight dots.

Figure 8 shows activity patterns due to four exemplar dots equally spaced from each other and from the location of a prototype dot. As the physical separation between the exemplar dots increases, the peaks of activity in the sum separate. At first, the distribution of activity in the sum as the dots separate simply seems to broaden the peak located at the prototype location. Then bumps due to the individual exemplars appear. Even with dots well separated, there is still substantial representation at the prototype location. It is clear that some representation of variability (the width of the activity pattern) is present as well as representation of the central tendency. In Figure 8, the left hand side shows the actual sum, the right hand side has the maximum values in the pattern drawn as the same height so relative curve shapes can be compared.

Figure 9 shows the formation of a sum from two exemplars (above) or from eight exemplars (below) In the diagrams, the exemplar dots have been chosen to be equally spaced about a central location corresponding to the dot in the prototype pattern. In both sums, the activity at the prototype location is substantial. In the two-exemplar sum, the activity is greatest at the location of the individual exemplars, but in the eight exemplar sum, the activity is largest at the prototype location, and the overall distribution of activity is more uniform. the smaller number of exemplars produces a "lumpy" sum dominated by learned exemplars, while the larger number yields a "smoother" sum with the prototype enhanced.

<u>The Computer Simulation of Prototype Extraction</u>. Our computer program is a very straightforward realization of the distributed model. Prototype dot patterns were generated randomly, with parameters identical to those used in the

experiments. Exemplars were formed according to the rules followed in the experiments. Sums of exemplars were constructed and a 'memory' was thereby formed. If there were three prototypes in a particular experiment, then three sums were constructed and kept separate. An input was classified as to which sum it was most similar to by computing the inner product with each sum in turn and choosing the largest. The sum with greatest similarity was the classification of the input, and the associated response was assumed to occur.

In the experiments described previously, and modeled in our simulations, we used groups containing 1, 6, and 24 exemplars of particular prototypes. It might be pointed out, with justice, that a prototype cannot be formed when only a single exemplar is presented. This is, of course, correct. However the model will generate similarities for new exemplars and the prototype for this case just as it will when many exemplars are presented so we should be able to predict the responses in this special case with the same model we used for multiple exemplars. A model for 'generalization' is to the prototype extraction model and should be predictable as a special condition of a prototype experiment.

The statistics of the patterns used in the simulations were identical in all respects to those in the real experiments. The only parameter to vary was the length constant of the fall off of the individual dot activity patterns. A total of 50 different sets of 24 patterns was used for each value of length constant. Thus a total of 50 different dot patterns were used for the one exemplar case, 300 for the 6 exemplar case, and 1,200 for the 24 exemplar case.

It is instructive to look at the qualitative results for different length constants. Figure 10 shows the results for four different values of length constant.

Figure 10 about here

Very short length constants produce a system which is powerfully biased toward response to old exemplars. There is relatively little prototype enhancement since the dot patterns are so widely separated that they do not interact. The old exemplars are recognized best, prototype next best, and new exemplars least well. As the length constant increases, the prototype is relatively enhanced and starts to generate a stronger response than the old exemplars. With very long length constants (very slow fall off of activity) responses tend to become more equal to all patterns, presumably because there is so little difference in activity patterns that there is little to discriminate among them.

The length constant found for best correlation in the similarity experiment was 14.5. Figure 11 shows the results when this value was used in the prototype simulation program. Although there is no way to make a direct mapping of inner product into percent correct response without numerous additional assumptions, there should be a monotonic relation between similarity and percent correct classification. The results in Figure 11 seem to show (to our biased eyes) remarkable qualitative similarity to the experimental results in Figure 6. Note in particular the crossing over of the responses to old exemplars shown for the six-exemplar case and the coming together of the responses to old and new exemplars in the 24 exemplar case. There is a strong response to the single old exemplar in the one-exemplar case, but the responses to the prototype and new exemplars for this (seemingly different) case fall right in line with the other data.

Figure 11 about here

Note that by changing length constant we can shift the emphasis of the system toward response to old exemplars at small length constants and toward prototype response at longer length constants. We investigate this next.

When is a Prototype Not Extracted (And Why). Almost as interesting a phenomenon as the enhancement of prototypes is when prototypes are not enhanced. We can predict when this occurs with our model. Very large distortions, which are judged not to be very similar might not give rise to much prototype enhancement. It is actually easy to see how this could occur. Let us consider the surfaces presented in Figure 8. Note that in large distortions, though there is enhancement at the prototype location, the peaks representing the individual exemplars are larger than the value of the sum at the prototype location. In the smaller distortions, there may be an enhancement of the prototype over the individual exemplars.

Let us analyze the simple case presented in Figure 8. Four exemplars are presented equally spaced from the prototype location (0,0), and from each other at locations (+$\underline{x}$,0), (0,+$\underline{x}$), (0,-$\underline{x}$) and (-$\underline{x}$,0). We assume the decay function is exp (-$\underline{r}$) where $\underline{r}$ is the distance from the peak, as before, with length constant 1. Let us consider only a single dot location. Two dots are separated from it by a distance $\sqrt{2}\ \underline{x}$. The dot on the other side of the square is 2 $\underline{x}$ distant. At the prototype location, (a distance $\underline{x}$ from each exemplar) the amplitude of the sum, $\underline{s}$, is given by

$$\underline{s}(0,0) = 4 \exp(-\underline{x}).$$

At the location of, say, the exemplar at (0, $\underline{x}$), the value of the sum is given by

$$\underline{s}\ (0, \underline{x}) = 1 + 2 \exp(-\sqrt{2}\ \underline{x}) + \exp(-2\ \underline{x}).$$

The sums at the other three dot locations are identical.

It can be shown that if $\underline{x}$ is less than 0.692 there is actual enhancement at the prototype location, that is the amplitude of the sum at (0,0) is greater than at any dot location.

The situation can be analyzed in more detail analytically, but without much improvement in our intuitive feeling for the system. in our experiments. We are interested in a more complex quantity, the relative sizes of the inner product between the prototype or an exemplar and the sum of old exemplars. This can be a relatively complicated function since the shape of the activity patterns due to a single dot are involved.

We can easily simulate this situation, however. Let assume that a number of exemplars are located equally spaced on the unit circle, with the prototype at the center of the circle. The number of exemplars in the simulations varied from two to a large number. We varied the length constant and computed inner products between an input pattern located at the prototype location and at the location of one of the dots on the circle. Figure 12 presents the ratio of the memory inner product at the prototype location and at the location of an exemplar.

Figure 12 about here

It can be seen that when the distance to the exemplars is very small (in length constants) the values at prototype and exemplar locations are almost identical, because everything adds without significant falloff. This case corresponds to very small average dot movement.

When the distance to the exemplars is very large (in length constants) activity from one exemplar decays to almost zero before it encounters activity from another exemplar. There is essentially no representation at the prototype location. This case corresponds to storage only of single exemplars with no prototype formation. This corresponds in the experiments to very large average movement of dots.

There is also a region of optimal prototype enhancement, which reaches a peak value around 20% greater than the value at the location of any individual exemplar. We have plotted on the

Figure 12 the length constant for best correlation, 14.5, derived from the similarity measures of Experiment 1.

Note that the actual experimental values we used were quite close to the point of maximum prototype enhancement. Although the theoretical justification for this was not available when the experiments were performed, it seems the intuitions and pilot studies of the experimenters were sound, since we desired to obtain the largest prototype effects possible.

Our experiment complements a similar investigation by Homa et al. (1973; see also Homa & Chambliss, 1975) in which the number of exemplars presented to subjects during category learning was varied from three to nine. On a subsequent classification test, subjects clasified all the training stimuli with near perfect accuracy. Performance on new exemplars and prototypes was worse that on old exemplars, but improved with size of learning set. Homa et al. used considerably greater distortions to generate the exemplars than we did; the experimental setups differ somewhat but we estimate the average dot displacement in their experiment to be about four length constants. This value of displacement would not be expected to yield prototype enhancement above response to old examples, which was the experimental finding. (At four space constant displacment, the maximum relative enhancment of the prototype is about 0.7). Also, Homa et al. found as the number of exemplars (corresponding to more dots spaced around the unit circle in the simulation) increased, the relative prototype enhancement increased, as would be predicted from the family of curves shown in Figure 12.

General Discussion

We have attempted to account for some aspects of human categorization with a model whose primary assumption holds that memory traces can lose their individuality when added together in storage. Although we are under no illusions about the neurophysiological rigor of the model as formulated here, the agreement obtained between theory and data suggests that a neurally inspired explanation of categorization at least deserves consideration along with rival models. However, two difficulties lie in the way of attempts to pit the distributed memory model against competing proposals. First, in order for the distributed memory model to generate specific predictions, one must make assumptions about how stimuli might be represented. We have done this for dot patterns, but not as yet for more commonly used stimuli. Second, it is notoriously difficult to get the various categorization models to make differential predictions (Hintzman & Ludlam, 1980; Smith & Medin, 1981, p. 182.) Indeed, several different models can be made to predict the experimental results reported here.

Consider a pure prototype model that stores only the average of learned instances, classifying novel stimuli according to their similarity to the average. Such a model predicts the results of Experiment 2 as follows. For the one item category, the stored average is simply the lone learned item. Naturally this pattern will be classified most accurately on a subsequent test. The objective prototype will have an advantage over other new patterns, being more similar to the learned pattern than a randomly selected exemplar (25.2 RMS units versus 38.5 RMS units). As category size increases, the pure prototype model predicts that classification of the prototype should improve, because the computed average matches the actual prototype more and more closely. Meanwhile, accuracy for the old and new exemplars should converge, because they are equally similar to the prototype.

Although it is possible to make differential predictions between the prototype and exemplar models, simply based on consideration of some special cases, the most telling differences arise when one considers the effectively dual representation of exemplars and prototype found in Figure 9, say. Here, the response to a dot at the prototype location will be identical for 2 exemplars presented four times each and for 8 exemplars presented once. The inner product in this case will depend only on the eight distances between center and exemplar and will not depend on exact arrangement of the exemplars. But the response to an old exemplar will be greatly different in the two cases: much larger for the small number of exemplars. A simple prototype model would not predict this because the average location is identical in both cases and the displacement from the average location of an old exemplar is the same.

Exemplar models, too, can be used to explain the present findings. As a trivial example, one can construct a model that stores learned exemplars separately, yet classifies test items using a similarity computation equivalent to the inner product function proposed here. It is difficult to conceive of a theoretical motivation for such a model, however, whereas the distributed memory categorizer is grounded in a theory of associative memory. One well motivated exemplar model that also predicts some of the present results is the context model (Medin & Schaffer, 1978). The context model computes the similarity between two items as the product of their similarity values along one or more featural dimensions. As does the inner product function, this multiplicative operation can judge two nearly identical stimuli as being dissimilar if they differ in one highly salient attribute. For the distributed memory model, the analogous situation occurs when corresponding elements in two activity patterns are large compared to other elements and have opposite signs, thus contributing negatively to the inner product. It is difficult to compare the two models further, because they assume such different forms of mental representation, but if the context model's featural dimensions are equated with individual dots in a pattern, and if similairty along one such dimension is assumed to vary with dot displacement, the the effect of displacement variance (Experiment 1) can be accounted for by the multiplicative rule.

The context model was prompted by Medin and Schaffer's finding that a test item's similarity to learned exemplars was a more important determinant of subjects' classification accuracy than the item's similarity to the category prototype (a result which incidentally argues against a pure prototype model). Since average old-new similarity in Experiment 2 was correlated with category size, the context model can explain the observed effects of manipulating the number of learned exemplars too. However it is not presently clear that old-new similarity is always a determining factor of accuracy in categorization tasks. Homa, Sterling, & Trepel (1981) trained subjects to categorize highly distorted exemplars and explicitly manipulated the similarity between old and new exemplars. Although classification accuracy for new patterns depended strongly on their similarity to old exemplars, the effect of old-new similarity was greater for small categories (5 old exemplars) than for large ones (20 old exemplars) a prediction of the distributed memory model. This effect was especially pronounced for the category prototypes, which were progressively enhanced in the larger categories (as also found in Experiment 2). This finding was not well predicted by the context model.

Homa et al. (1981) interpreted their results as support for a mixed model of categorization in which a category's representation includes both abstracted prototypes and individual exemplars, with their relative weights determined by a number of aspects of the experimental situation. It is worth pointing out that two of these aspects, category size and within-category variability, are the primary determinants of exemplar versus prototype dominance in the distributed memory model. Perhaps the distributed memory model may best be considered as a particular embodiment of the mixed model proposed by Homa et al. (1981), since it both stores instances and given suitable input, abstracts a prototype. Furthermore, the effects of category size and variability on the nature of a category's representation find straightforward explanations in the distributed memory model, because those effects are direct consequences of the proposed memory storage operation.

The mixed behavior of the distributed memory model deserves an additional comment. Intuition tells us that different categories demand different kinds of mental representations. It seems inefficient to have always to test a novel item against a large number of highly similar traces, as required by a strict exemplar view. Conversely, representing a small set of diverse category members by their average, as required by a pure prototype view, invites errors. An attractive feature of the distributed memory model is that it automatically tailors its representation of a category to accomodate the variability of the learned items.

Let us take an important example from outside the world of dot patterns. Intuitively, we feel some things are too different from each other to form 'natural' equivalence classes, though it is always possible to explicitly bind very different things together to form complex concepts. As an example, consider 'A' and 'a'. In some contexts, these two forms represent the same noises and partake of identical interpretations. In other contexts, they are clearly separated, having different names such as 'Capital A' and 'small a.' It seems to us highly unlikely and even rather undesirable that presentation of 'A' and 'a' and distortions of them would lead to formation of their average as the best example of the written letter 'ay'. In this case, it seems best to assume that there are two simple prototypes formed which in some contexts are given the same names. Formation of a complicated concept from two simple ones is quite straightforward in the association model, as a matter of fact. Suppose $\mathbf{f}_1$ and $\mathbf{f}_2$ are given the same name, $\mathbf{g}$, and $\mathbf{f}_1$ and $\mathbf{f}_2$ are sufficiently different to be orthogonal to one another. Then we form the overall $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{g}\,\mathbf{f}_1^T + \mathbf{g}\,\mathbf{f}_2^T .$$

Presentation of either $\mathbf{f}_1$ or $\mathbf{f}_2$ would evoke the correct output, $\mathbf{g}$. If it were possible to present the sum, $(\mathbf{f}_1 + \mathbf{f}_2)$ we would also evoke the correct output, with larger amplitude. Sometimes it is not possible to present the two items simultaneously: suppose 'A' give rise to coding $\mathbf{f}$ and 'a' gives rise to coding $\mathbf{f}$. Presentation of 'A' superimposed on 'a' does not generally give the sum $\mathbf{f}_1 + \mathbf{f}_2$ as the resulting neural representation because of a number of kinds of well understood effects in the initial stages of the visual system. If, however, the letters are presented in sequence or spatially separated, generating something like a sum with less mutual interference, it seems intituitive that the name 'ay' is indeed likely to be evoked as the strongest common association of two different state vectors. We can use this technique computationally to extract common associations from ambiguous or complicated concepts, (Hayes-Roth & Hayes-Roth, 1977; Elio & J.R.Anderson, 1981) an application to be described elsewhere.

Clearly, additional theoretical and experimental work will be required to differentiate the distributed memory model from other models of categorization. Even without definitive experimental validation, the approach we have outlined can draw support from the following observations: (1) The same model has been successfully applied to a variety of other cognitive behaviors; (2) The model is robust with respect to its most tenous assumptions, those concerning the details of the representation of stimuli; (3) The model correctly predicts the quantitative (Experiment 1) and qualitative (Experiment 2) results of the experiments presented here, and does so using a single value of its one adjustable parameter. We hope this discussion raises the possibility that theoretical investigations of neural processes may shed light on psychological theory.

References

Abramowitz, M. & Stegun, I.A. (Eds.) Handbook of Mathematical Functions. National Bureau of Standards Applied Mathematics Series: Number 55. Washington, D.C.: U.S. Government Printing Office, 1964.

Anderson, J.A. A theory for the recognition of items from short memorized lists. Psychological Review, 1973, 80, 417-438.

Anderson, J.A. Neural models with cognitive implications. In D. LaBerge & S.J. Samuels (Eds.), Basic Processes in Reading. Hillsdale, N.J.: Erlbaum Associates, 1977.

Anderson, J.A. Neural models for cognition. Proceedings of the I.E.E.E.: System, Man, and Cybernetics, (in press)

Anderson, J.A. & Hinton, G.E. Models of information processing in the brain. In G.E. Hinton & J.A. Anderson (Eds.), Parallel Models of Associative Memory. Hillsdale, N.J.: Erlbaum Associates, 1981.

Anderson, J.A., Silverstein, J.W., Ritz, S.A., & Jones, R.S. Distinctive features, categorical perception, and probability learning: some applications of a neural model. Psychological Review, 1977,84, 413-451.

Barresi, J., Robbins, D., & Shain, K. Role of distinctive features in the abstraction of related concepts. Journal of Experimental Psychology: Human Learning and Memory, 1975, 104, 360-368.

Bransford, J.D. & Franks, J.J. The abstraction of linguistic ideas. Cognitive Psychology, 1971, 2, 331-350.

Brooks, L. Non-analytical concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds.), Cognition and Categorization. Hillsdale, N.J.: Erlbaum Associates, 1978.

Collins, A.M. & Quillian, M.R. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 240-248.

Eich, J.M. A composite holographic associative recall model. Psychological Review, 1982, 89, 627-661.

Elio, R. & Anderson, J.R. The effects of category generalizatons and instance similarity on on schema abstraction. Journal of Experimental Psychology: Human Learning and Memory, 1981, 7, 397-417.

Franks, J.J. & Bransford, J.D., Abstraction of visual patterns. Journal of Experimental Psychology, 1971, 90, 65-74.

Ghiselin, M.T. Categories, life, and thinking. The Behavioral and Brain Sciences, 1981, 4, 269-313.

Goldman, D. & Homa, D. Integrative and metric properties of abstracted information as a function of category discriminability, instance variability, and experience. Journal of Experimental Psychology: Human Learning and Memory, 1977,3,18-28.

Gradshteyn, I.S. & Ryzhik, I.M. Table of Integrals, Series, and Products, Fourth Edition. New York: Academic Press, 1965.

Hartley, J. & Homa, D. Abstraction of stylistic concepts. Journal of Experimental Psychology: Human Learning and Memory, 1981, 7, 33-46.

Hayes-Roth, B. & Hayes-Roth, F. Concept learning and the recognition and classification of exemplars. Journal of Verbal Learning and Verbal Behavior, 1977, 16, 321-338.

Hebb, D.O. The Organization of Behavior. New York: Wiley, 1949.

Hinton, G.E. & Anderson, J.A. (Eds.). Parallel Models of Associative Memory. Hillsdale, N.J.: Erlbaum Associates, 1981.

Hintzman, D.L. & Ludlam, G. Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. Memory and Cognition, 1980, 8, 378-382.

Homa, D. & Chambliss, D. The relative contributions of common and distinctive information on the abstraction from ill-defined categories. Journal of Experimental Psychology: Human Learning and Memory, 1975,1, 351-59.

Homa, D., Cross, J., Cornell, D., Goldman, D. & Shwartz, S. Prototype abstraction and classification of new instances as a function of the number of instances defining the prototype. Journal of Experimental Psychology, 1973, 101, 116-122.

Homa, D., Sterling, S., & Trepel, L. Limitations of exemplar-based generalization and the abstraction of categorical information. Journal of Experimental Psychology: Human Learning and Memory, 1981, 7, 418-439.

Kohonen, T. Correlation matrix memories. IEEE Transactions on Computers, 1972, C-21, 353-359.

Kohonen, T. Associative Memory: A System Theoretic Approach. Berlin: Springer-Verlag, 1977.

Levy, W., Anderson, J.A., & Lehmkuhle, W. (Eds.). Synaptic Change in the Nervous System. Hillsdale, N.J.: Erlbaum Associates, 1977.

Korn, G.A. & Korn, T.M. *Mathematical Handbook for Scientists and Engineers, Second Edition*. New York: McGraw-Hill, 1968.

McClelland, J.L. & Rumelhart, D.E. An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review*, 1981, *88*, 375-497.

Medin, D.L. & Schaffer, M.M. Context theory of classification learning. *Psychological Review*, 1978, *85*, 207-238.

Mervis, C.B. & Rosch, E. Categorization of natural objects. *Annual Review of Psychology*, 1981, *32*, 89-115.

Millward, R.B., Aikin, J. & Wickens, T.D. The Human Learning Laboratory at Brown University. In *Computers in the Psychological Laboratory, Vol 2*. Maynard, Mass.: Digital Equipment Corporation, 1972.

Murdock, B.B., Jr. A theory for the storage and retrieval of item and associative information. *Psychological Review*, 1982, *89*, 609-626.

Neisser, U. *Cognitive Psychology*. New York: Appleton-Century-Crofts, 1967.

Nelson, K. Concept, word and sentences: Interrelations in acquisition and development. *Psychological Review*, 1974, *81*, 267-285.

Omohundro, J. & Homa, D. Search for abstracted information. *American Journal of Psychology,* 1981, *9*, 324-331.

Posner, M.I. Abstraction and the process of recognition. In J.T. Spence & G.H. Bower (Eds.), *Advances in Learning and Motivation (Volume 3)*. New York: Academic Press, 1969.

Posner, M.I., Goldsmith, R. & Welton, K.E. Perceived distance and the classification of distorted dot patterns. *Journal of Experimental Psychology*, 1967, *73*, 28-38.

Posner, M.I. & Keele, S.W. On the genesis of abstract ideas. *Journal of Experimental Psychology*, 1968, *77*, 353-363.

Posner, M.I. & Keele, S.W. Retention of abstract ideas. *Journal of Experimental Psychology*, 1970, *83*, 304-308.

Reed, S.K. Pattern recognition and categorization. *Cognitive Psychology*, 1972, *3*, 382-407.

Reed, S.K. Category V. Item learning: implications for categorization models. *Memory and Cognition*, 1978, *6*, 612-620.

Reitman, J.S. & Bower, G.H. Storage and later recognition of exemplars of concepts. Cognitive Psychology, 1973, 4, 194-206.

Rips, L.J., Shoben, E.J. & Smith, E.E. Semantic distance and the verification of semantic relations. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 1-20.

Robbins, D., Barresi, J., Compton, P., Furst, A., Russo, M. & Smith, M.A. The genesis and use of exemplar vs. prototype knowledge in abstract category learning. Memory and Cognition, 1978, 6, 473-480.

Rumelhart, D.E. & McClelland, J.L. An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. Psychological Review, 1982, 89, 60-94.

Smith, E.E. & Medin, D.L. The Psychology of Conceptual Processes, Cambridge, Mass.: Harvard University Press, 1981.

Stevens, K.A. Computation of locally parallel structure. Biological Cybernetics, 1978, 29, 19-28.

Strange, W., Keeney, T., Kessel, F.S. & Jenkins, J.J. Abstraction over time from distortions of random dot patterns -- a replication. Journal of Experimental Psychology, 1970, 83, 508-510.

White, B.W. Recognition of familiar characteristics under an unfamiliar transformation. Perceptual and Motor Skills, 1962, 15. 107-116.

Figure Captions

Figure 1. Models assume two sets of N, neurons alpha projecting to beta. Every neuron in alpha projects to every neuron in beta. This drawing has N = 6. From Anderson, Silverstein, Ritz and Jones, 1977. Reprinted by permission.

Figure 2. A prototype dot pattern ('P') followed by five examples at various levels of distortion. Dots were generated on a 512 by 512 array and presented to subjects on a CRT screen. The number refers to the average number of locations moved on the array. A distance of 100 locations is indicated.

Figure 3. Mean similarity ratings as a function of average displacement. Filled circles, no-variance pairs, open circles high-variance pairs. The horizontal variablity of the high variance points is smaller than the width of the symbols.

Figure 4. Activity pattern on a hypothetical cortex due to a single dot in the real world. There is an exponential decay with distance from a central point. Height corresponds to activity. Length constant of the exponential is shown.

Figure 5. Value of integral between two activity patterns with exponential falloff and relative displacement of their centers. The graph is presented with the y axis inverted to correspond to the graph of experimental data presented in Figure 3.

Figure 6. Percentage of correct test trial classifications for Experiment 2.

Figure 7. Correct reaction times for classification for Experiment 2.

Figure 8. The sum of four exemplars each due to a single dot. Demonstration of the effect of increasing spacing between dots. There is relatively greater representation of individual exemplars as spacing increases but still a significant buildup at the prototype location. The left hand column gives the simple sum of the individual exemplar activity patterns; the right hand column equates the maximum activity of each displacment so shapes of curves can be compared.

Figure 9. Demonstration of the qualitative difference between a 'memory' sum constructed from two different exemplars and one constructed from eight different exemplars. Exemplars in both cases had the same average separation from the prototype location. Note the very 'lumpy' memory produced when only a few exemplars are stored. In this case, there is representation of item information; when many exemplars are stored, the prototype gives the largest response.

Figure 10. Simulations of Experiment 2 with various length constants, showing the shift from learning of specifics about old exemplars (small length constants) to prototype representation (with more presented exemplars and larger length constants).

Figure 11. Simulation of Experiment 2 using the best fitting length constant determined in the similarity experiment, Experiment 1. Compare with Figure 6.

Figure 12. This simulation asssumed varying numbers of dots (from 2 to 24) were arranged on the unit circle. The prototype was located at the center of the circle. The graph gives the ratio of the system response to a dot at the prototype location and the system response to a dot at the location of an old exemplar. The _Y_ axis gives the radius of the circle in length constants. The average displacement of patterns used in Experiment 2 is indicated on the Figure.

Table 1

Displacements used to create the distortions of

Experiment 1 and two distance measures

Level	Displacements	Average Distance	RMS Distance
0	0	0.0	0.0
15	15	15.0	15.0
30	30	30.0	30.0
45	45	45.0	45.0
60	60	60.0	60.0
90	90	90.0	90.0
15H	0-5, 25-30[a]	13.6 ± 0.5[b]	18.5 ± 0.5[b]
30H	0-10, 50-60	27.2 ± 1.0	36.8 ± 1.0
45H	0-15, 75-90	40.7 ± 1.8	55.4 ± 1.6
60H	0-20, 100-120	54.4 ± 2.1	73.7 ± 1.6
90H	0-30, 150-180	81.4 ± 2.8	111.1 ± 2.6

[a] Five dots were displaced by sampling uniformly from the first interval, four by sampling from the second interval.

[b] Mean ± standard deviation of 200 randomly chosen patterns at each level.

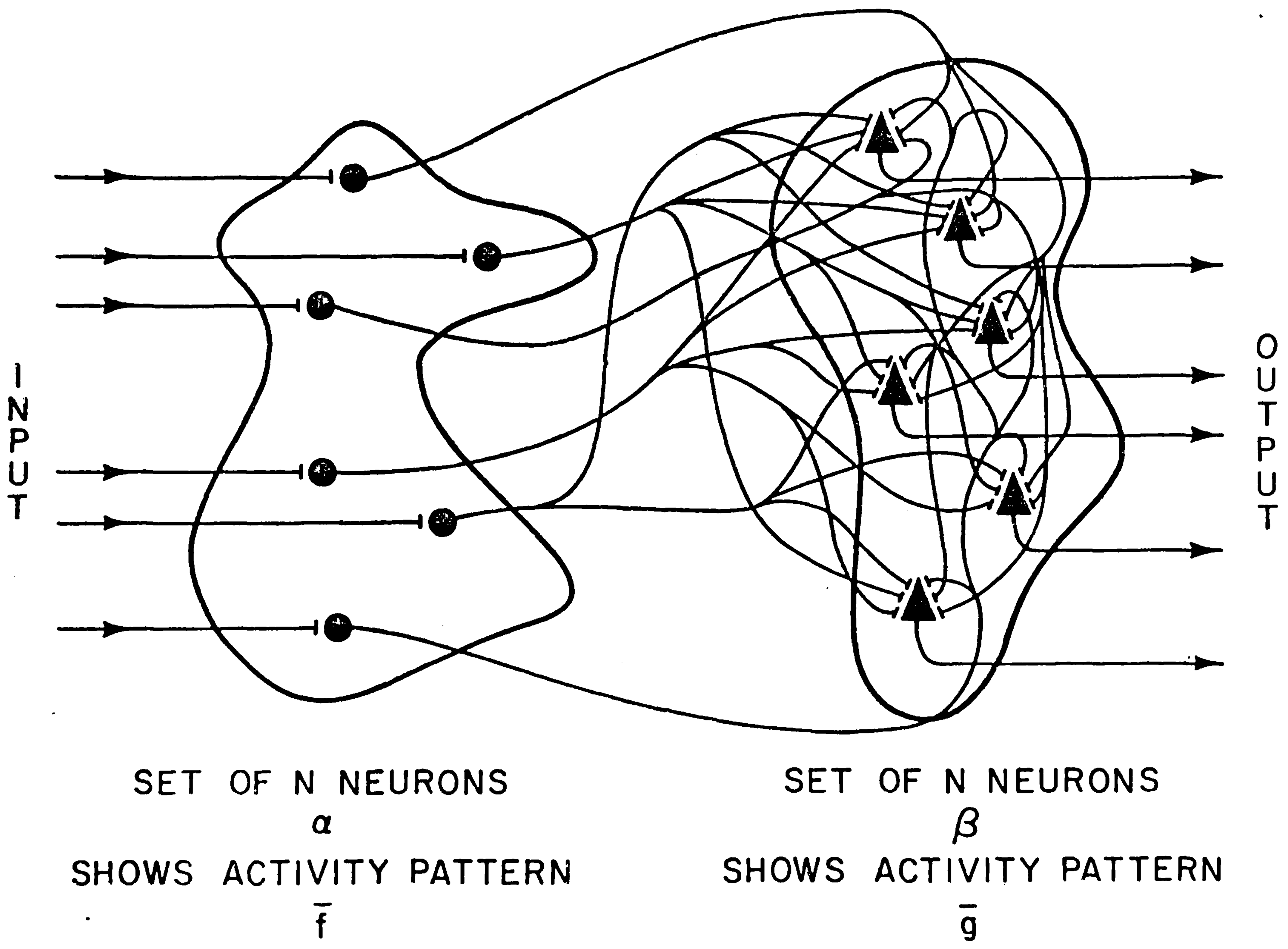
Table 2

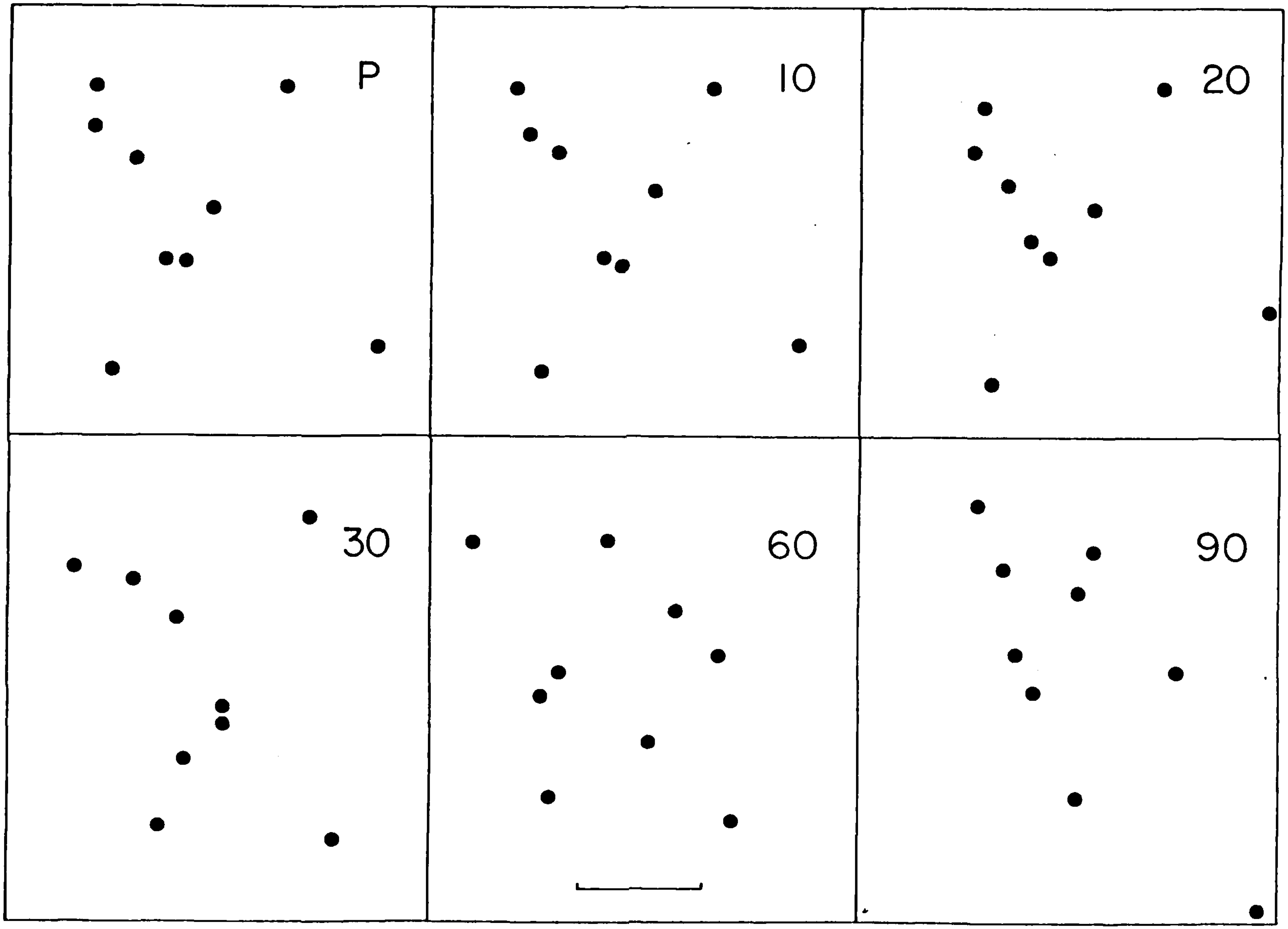
Experimental and computed similarity values for

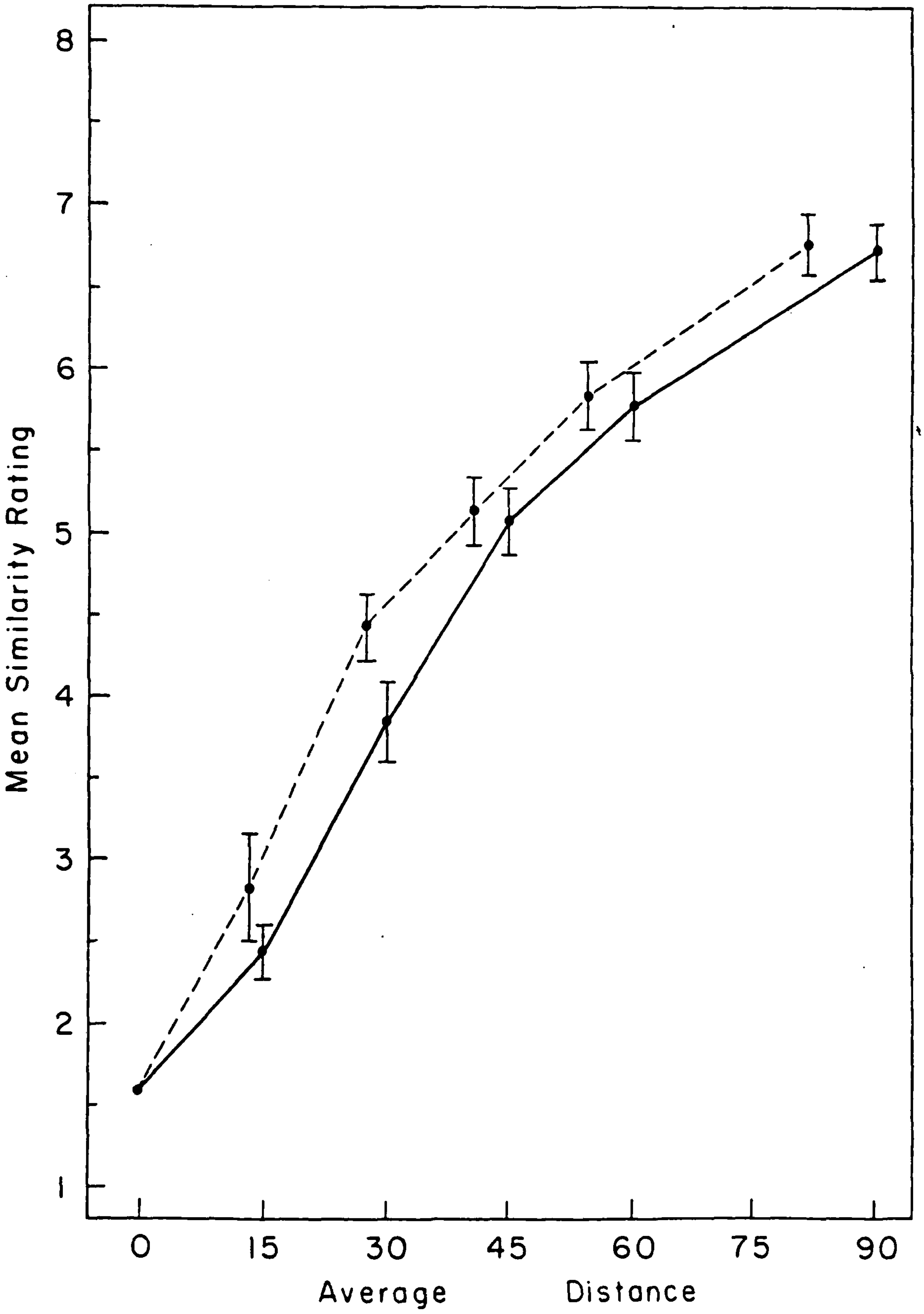
Experiment 1. Calculations as described in text

Level	Average Displacement in Length Constants	Observed Similarity	Computed Similarity
0	0.00	1.59	1.59
15	0.81	2.43	2.45
30	1.63	3.85	4.05
45	2.44	5.09	5.48
60	3.26	5.79	6.50
90	4.89	6.73	7.54
15H	0.74	2.82	2.43
30H	1.46	4.45	3.70
45H	2.21	5.16	4.64
60H	2.94	5.84	5.32
90H	4.42	6.76	6.23

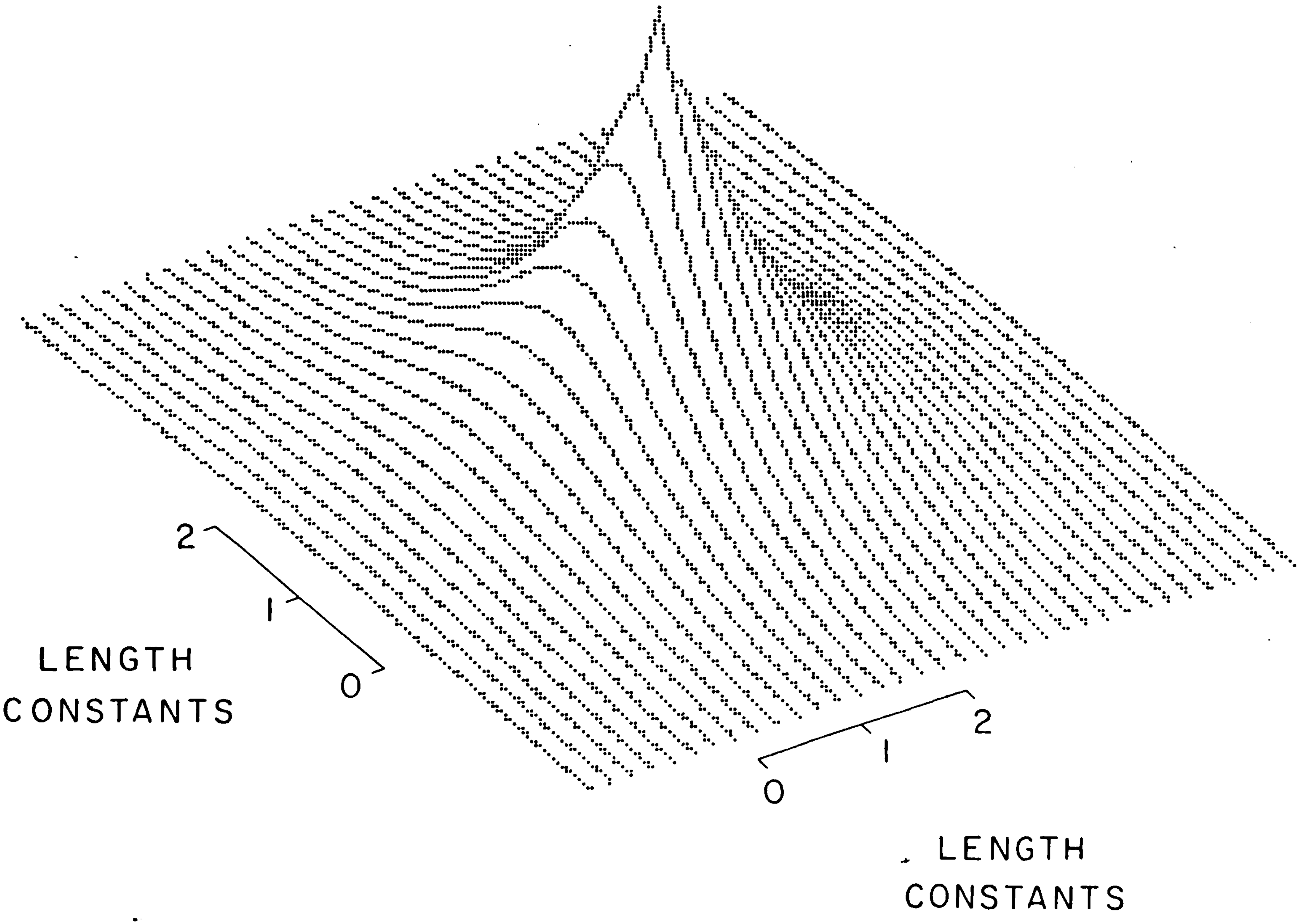
Note. The computed space constant for best fit was 18.42 oscilloscope screen units.

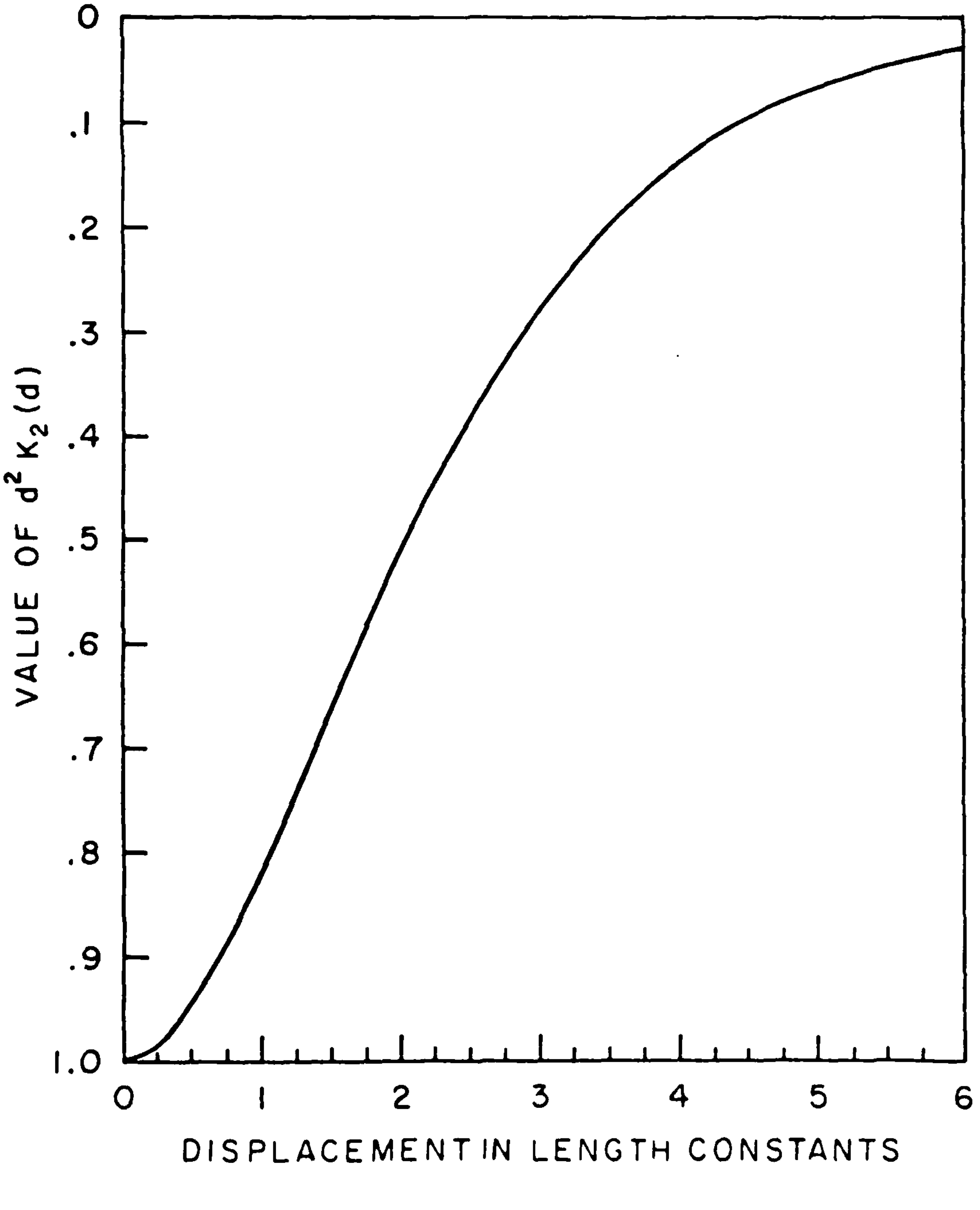

INPUT
OUTPUT
SET OF N NEURONS
α
SHOWS ACTIVITY PATTERN
$\bar{f}$
SET OF N NEURONS
β
SHOWS ACTIVITY PATTERN
$\bar{g}$

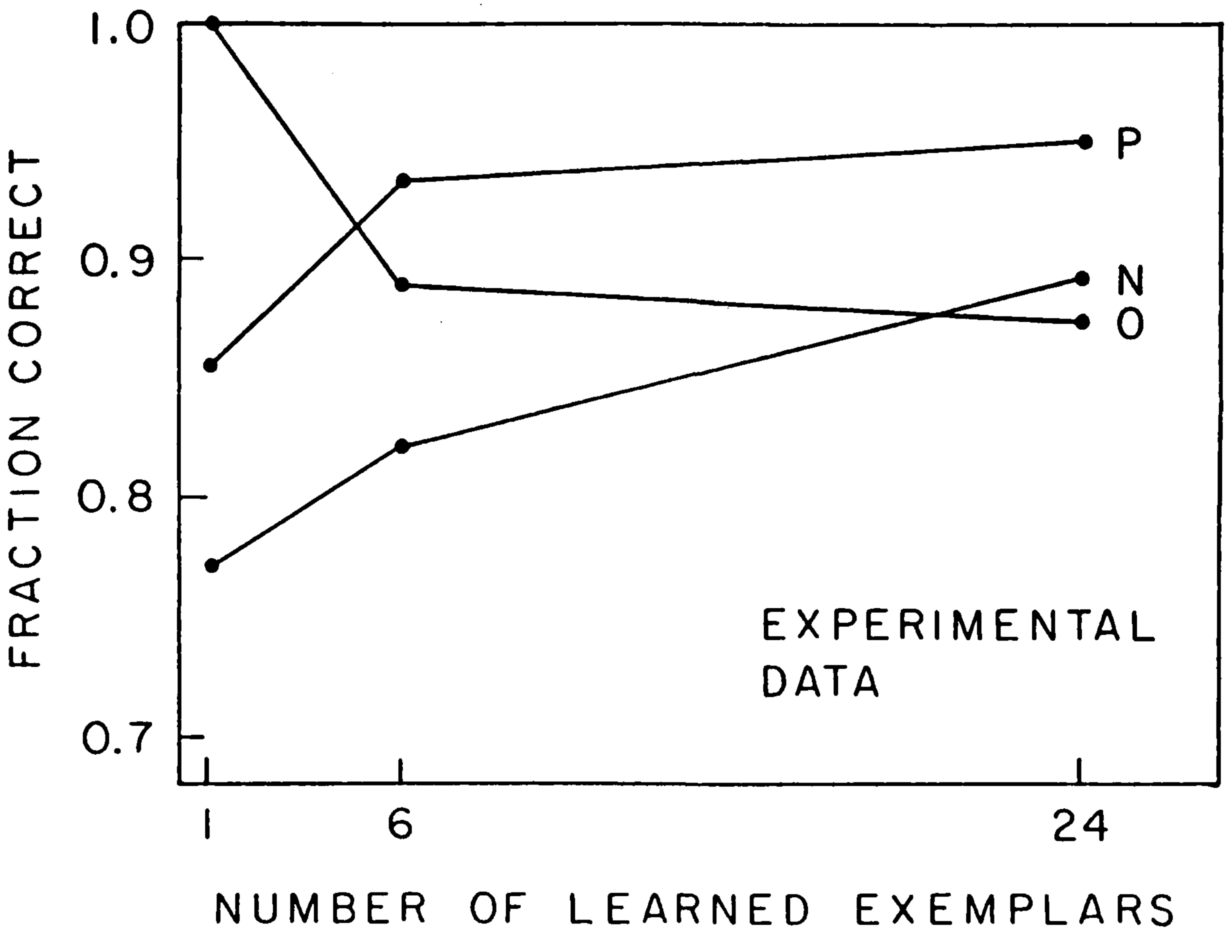

P
10
20
30
60
90

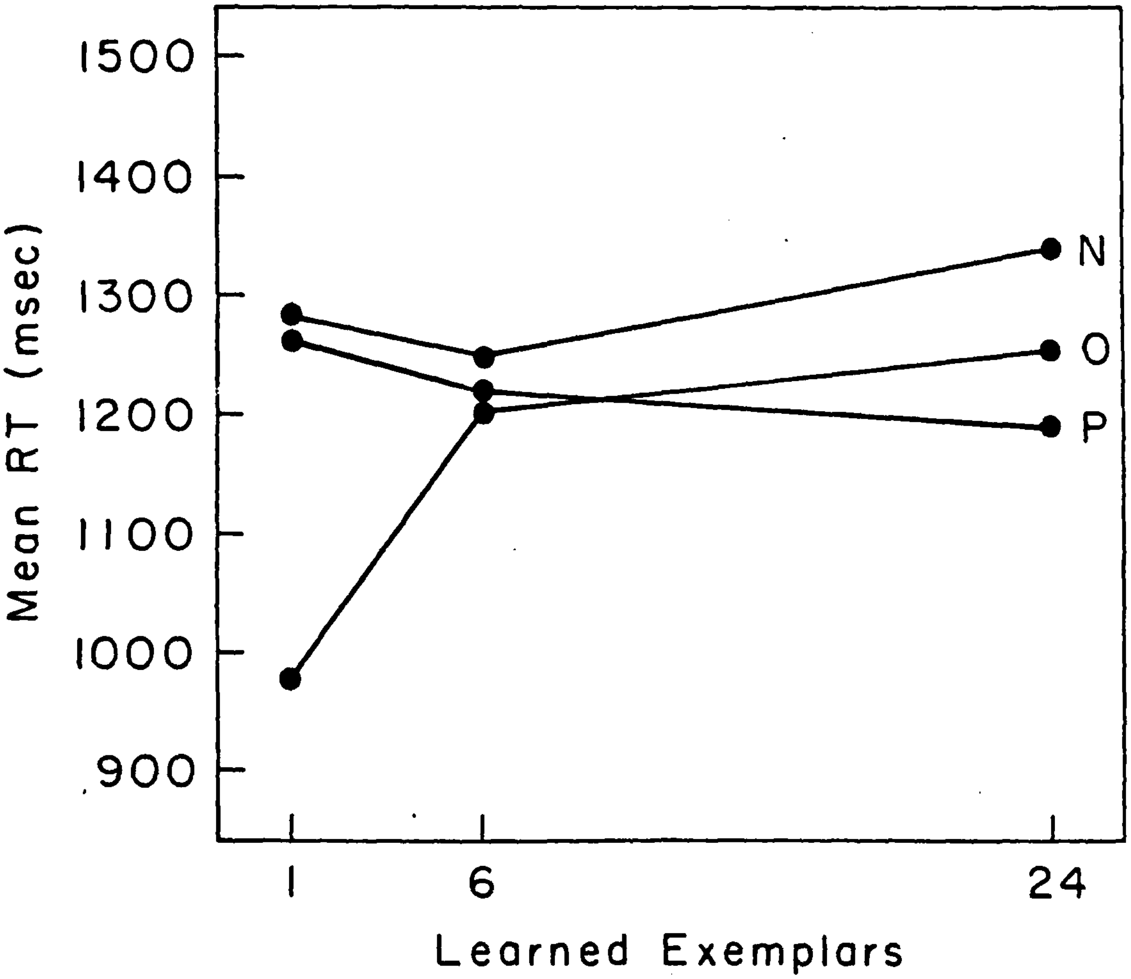

8
7
6
5
4
3
2
1
Mean Similarity Rating
0
15
30
45
60
75
90
Average Distance

RESPONSE TO A DOT

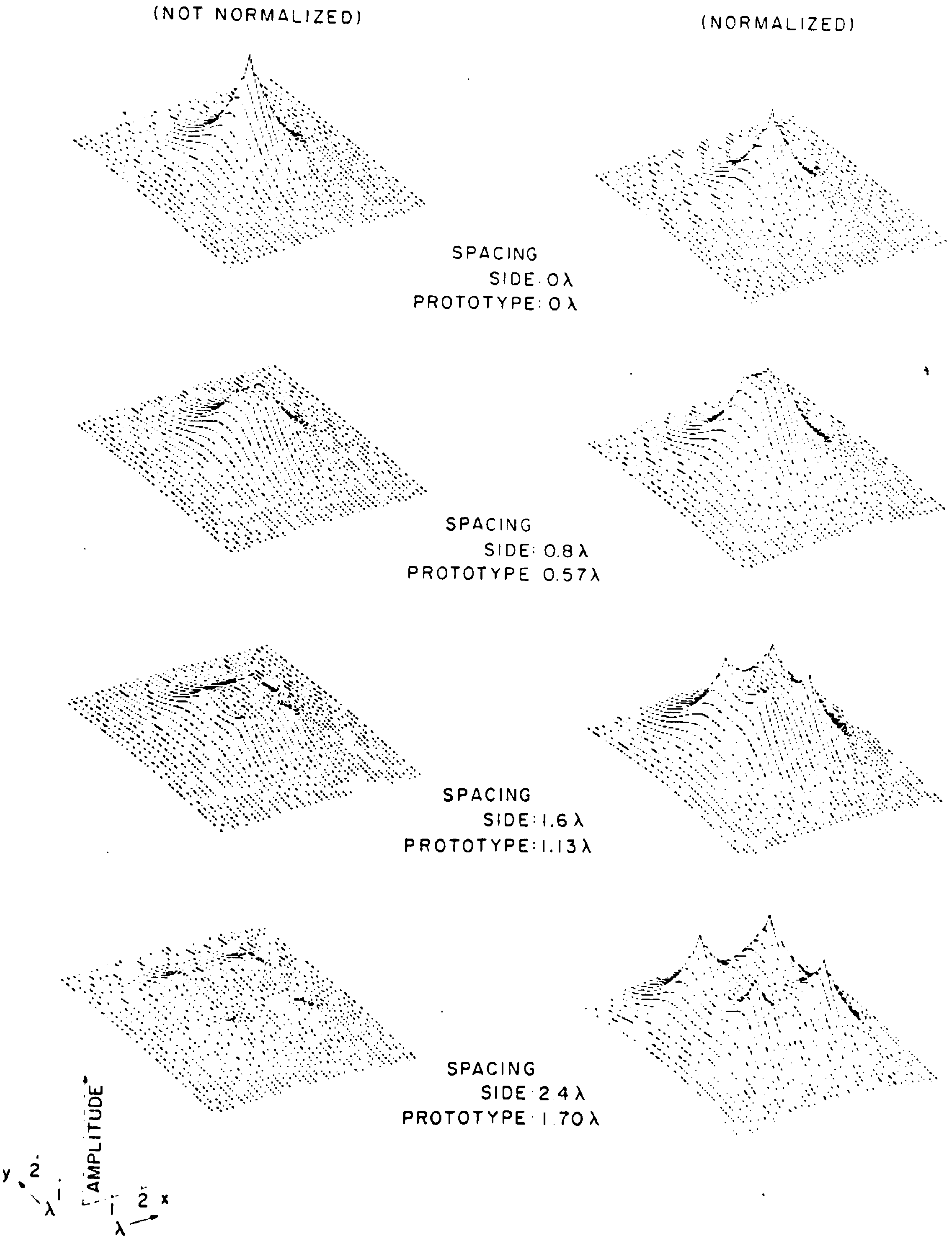



0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
VALUE OF $d^2 K_2(d)$
0
1
2
3
4
5
6
DISPLACEMENT IN LENGTH CONSTANTS

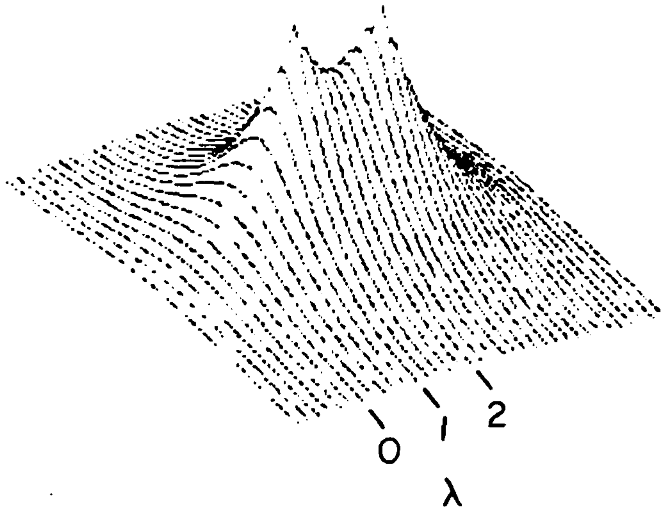

1.0
0.9
0.8
0.7
FRACTION CORRECT
1
6
24
NUMBER OF LEARNED EXEMPLARS
P
N
O
EXPERIMENTAL
DATA



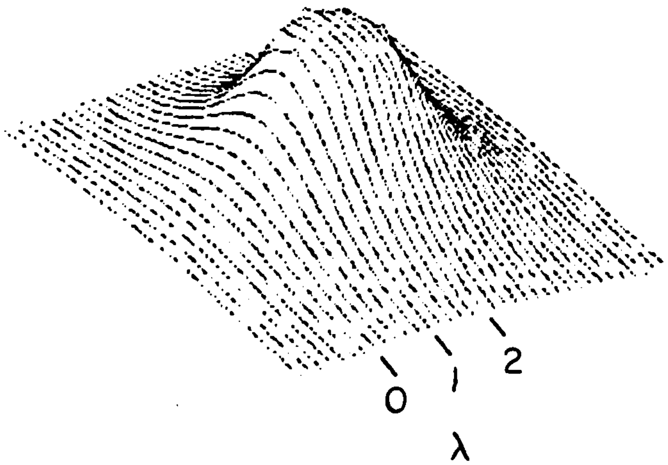
Mean RT (msec)
1500
1400
1300
1200
1100
1000
900
1
6
24
Learned Exemplars
N
O
P


(NOT NORMALIZED)
(NORMALIZED)
SPACING
SIDE: 0λ
PROTOTYPE: 0λ
SPACING
SIDE: 0.8λ
PROTOTYPE 0.57λ
SPACING
SIDE: 1.6λ
PROTOTYPE: 1.13λ
SPACING
SIDE 2.4λ
PROTOTYPE 1.70λ
AMPLITUDE
y
2
1
λ
1
2
x
λ

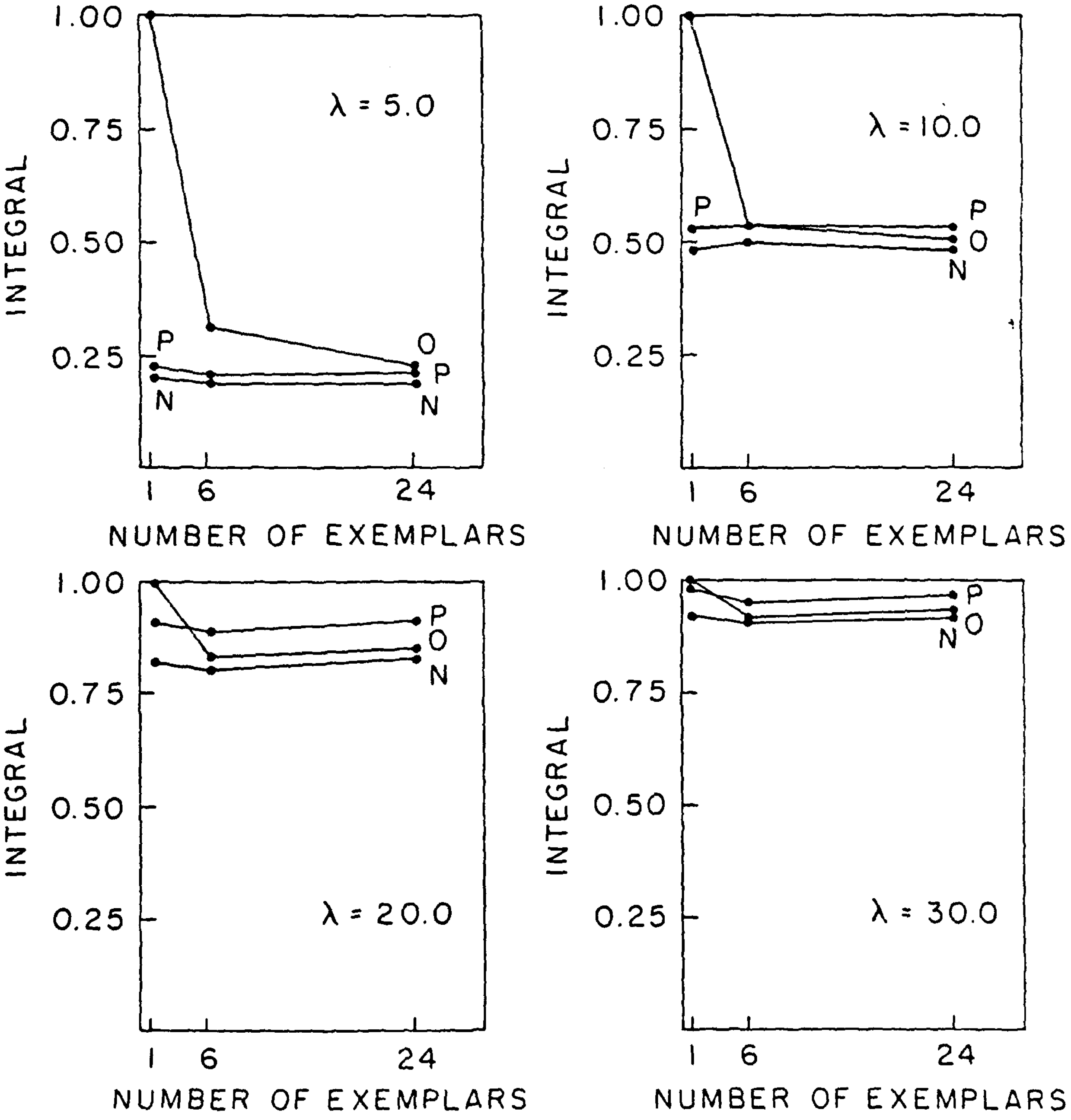
MEMORY FORMATION

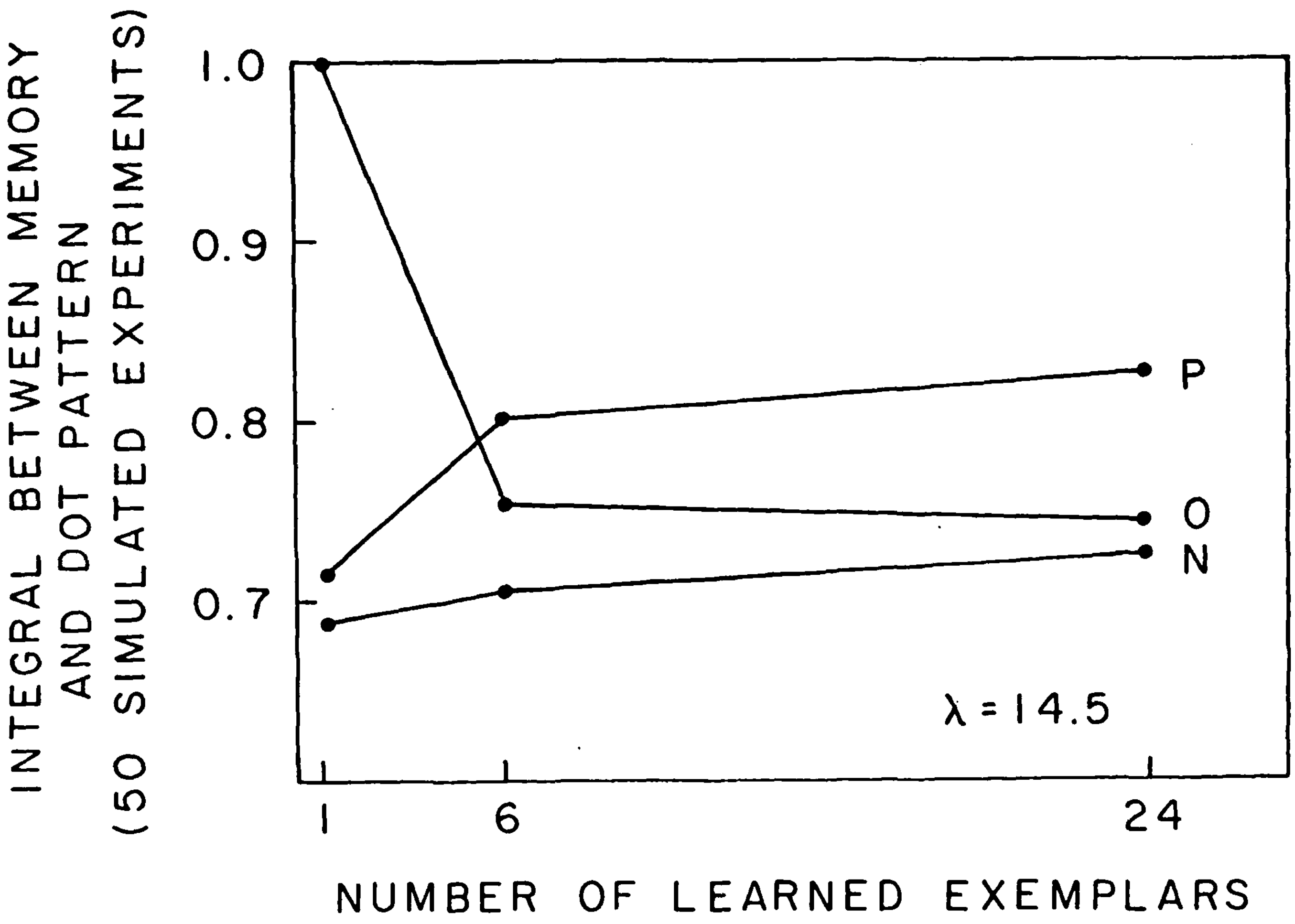


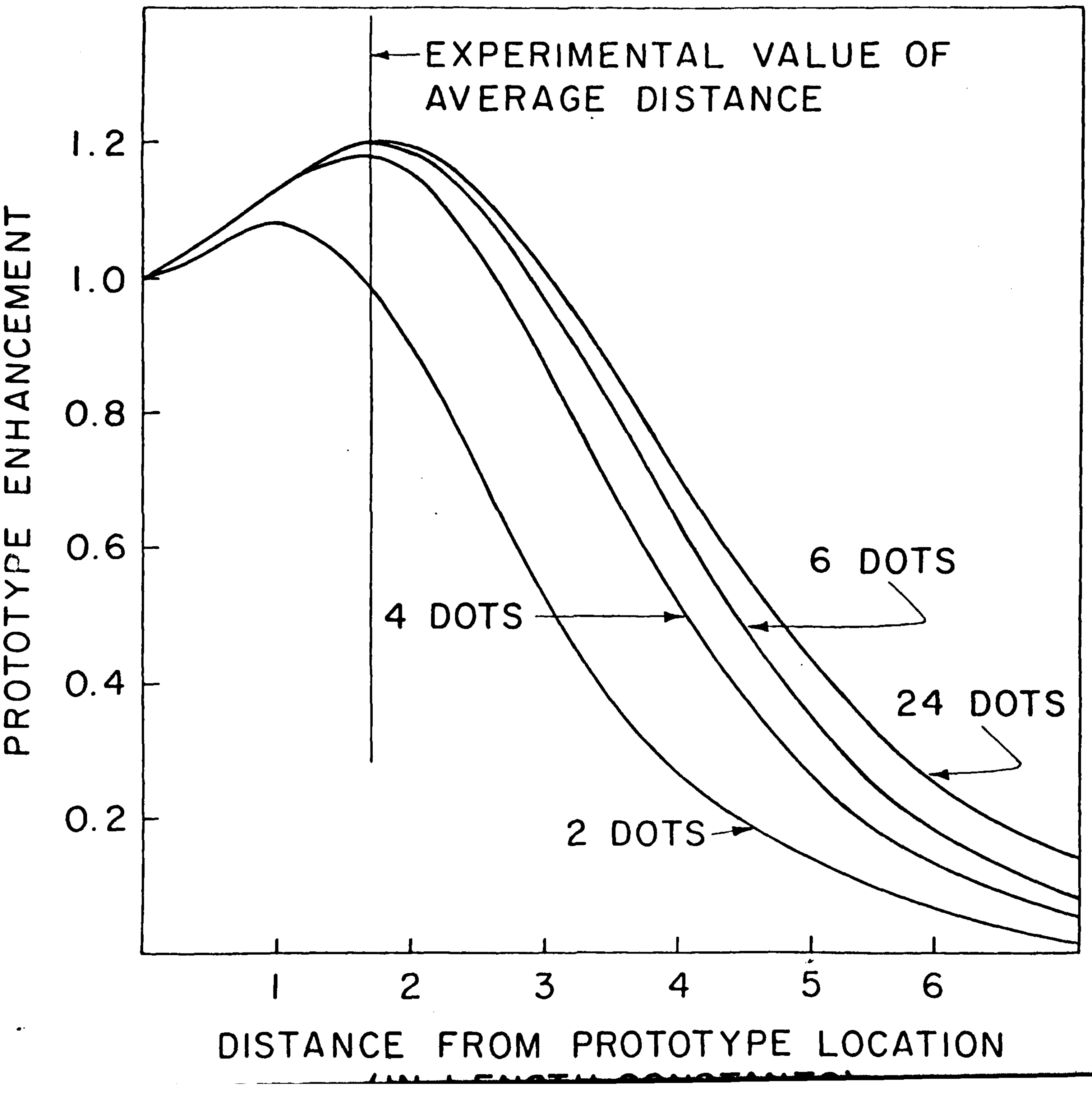
2 EXEMPLARS
4 PRESENTATIONS EACH



8 EXEMPLARS
1 PRESENTATION EACH


λ = 5.0
λ = 10.0
λ = 20.0
λ = 30.0
INTEGRAL
1.00
0.75
0.50
0.25
NUMBER OF EXEMPLARS
1
6
24
P
O
N


INTEGRAL BETWEEN MEMORY
AND DOT PATTERN
(50 SIMULATED EXPERIMENTS)
1.0
0.9
0.8
0.7
P
O
N
λ = 14.5
1
6
24
NUMBER OF LEARNED EXEMPLARS


EXPERIMENTAL VALUE OF
AVERAGE DISTANCE
PROTOTYPE ENHANCEMENT
1.2
1.0
0.8
0.6
0.4
0.2
6 DOTS
4 DOTS
24 DOTS
2 DOTS
1
2
3
4
5
6
DISTANCE FROM PROTOTYPE LOCATION
(IN LENGTH CONSTANTS)

FIL